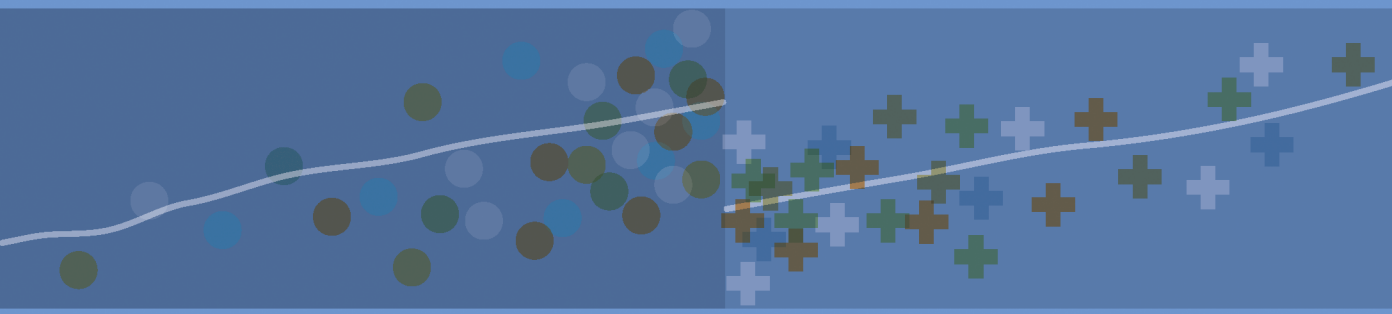
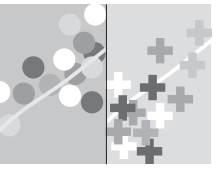
Avaliação de Impacto na Prática



Paul J. Gertler, Sebastian Martinez,
Patrick Premand, Laura B. Rawlings,
Christel M. J. Vermeersch

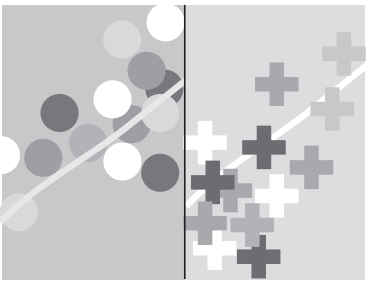
GRUPO BANCO MUNDIAL

Avaliação de Impacto na Prática



Este livro só foi possível graças ao generoso apoio do Fundo Estratégico para a Avaliação de Impacto (SIEF). Lançado em 2007, com uma doação de $14,9 milhões feita pela Espanha e expandido por uma doação de $2,1milhões do Departamento para o Desenvolvimento Internacional do Reino Unido (DfID), o SIEF é o maior fundo fiduciário voltado para a avaliação de impacto já estabelecido no Banco Mundial. Seu principal objetivo é expandir a base de evidências sobre o que funciona para melhorar os resultados na saúde, educação e proteção social, informando, desse modo, a política de desenvolvimento. Vide **http://www.worldbank.org/sief.**

Avaliação de Impacto na Prática



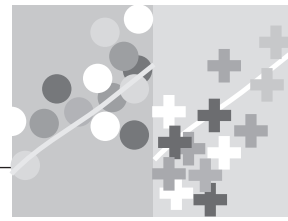
Paul J. Gertler, Sebastian Martinez,
Patrick Premand, Laura B. Rawlings,
Christel M. J. Vermeersch

GRUPO BANCO MUNDIAL

Este trabalho foi publicado originalmente em ingles pelo Banco Mundial como *Impact Evaluation in Practice* in 2011. Em caso de disprepancias, predomina o idioma original.

ISBN (paper): 978-1-4648-0088-7
ISBN (electronic): 978-1-4648-0090-0
DOI: 10.1596/978-1-4648-0088-7

Projeto da capa por Naylor Design.

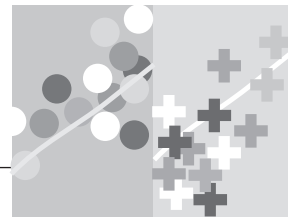


ÍNDICE

Caixas

Figuras

Tabelas



PREFÁCIO

Este livro oferece uma introdução acessível sobre o tema da avaliação de impacto e a sua prática no desenvolvimento socioeconômico. Embora o livro seja direcionado principalmente aos profissionais de desenvolvimento e gestores de políticas públicas, acreditamos que será um recurso valioso para estudantes e outros interessados no tema da avaliação de impacto. As avaliações de impacto prospectivas avaliam se um programa atingiu ou não os resultados pretendidos ou testam estratégias alternativas para atingir estes resultados. Consideramos que mais e melhores avaliações de impacto ajudarão a fortalecer a base de evidência para políticas e programas voltados para o desenvolvimento em todo o mundo. A nossa esperança é de que se os governos e profissionais de desenvolvimento puderem tomar decisões de políticas com base em evidências, incluindo evidências geradas por meio de avaliações de impacto – os recursos destinados às políticas de desenvolvimento serão gastos de modo mais efetivo para reduzir a pobreza e melhorar as vidas das pessoas. As três partes deste manual fornecem uma introdução não técnica à avaliação de impacto, discutindo, na primeira parte, o que avaliar e por que, na segunda parte discute-se como avaliar e, na terceira parte, como implementar uma avaliação. Estes elementos são as ferramentas básicas necessárias para realizar uma avaliação de impacto com sucesso.

A abordagem à avaliação de impacto neste livro é, em boa medida, intuitiva e tentamos minimizar a notação técnica. Oferecemos ao leitor um conjunto básico de ferramentas para a avaliação de impacto - os conceitos e os métodos que fundamentam qualquer avaliação de impacto - e discutimos a sua aplicação nas ações de desenvolvimento no mundo real. Os métodos são extraídos diretamente da pesquisa aplicada em ciências sociais e compartilham aspectos com os métodos de pesquisa usados nas ciências naturais. Neste sentido, a avaliação de impacto reúne as ferramentas de pesquisa empírica amplamente usadas na economia e outras ciências sociais, com as

realidades operacionais e de economia política envolvidas na implementação de políticas públicas e na prática do desenvolvimento.

De um ponto de vista metodológico, nossa abordagem de avaliação de impacto é, em grande medida, pragmática: achamos que os métodos mais apropriados devem ser identificados de modo a corresponderem ao contexto operacional, e não o contrário. Isso é melhor alcançado no início de um programa, com o desenho de avaliações de impacto prospectivas como parte integrante da implementação do projeto. Argumentamos que a obtenção de consenso entre as principais partes interessadas e a identificação de um método de avaliação que se adapte ao contexto político e operacional são aspectos tão importantes quanto o método em si. Também acreditamos que as avaliações de impacto devam ser honestas quanto as suas limitações e condicionantes. Finalmente, encorajamos fortemente os formuladores de políticas e gestores de programas a considerar as avaliações de impacto numa estrutura lógica que defina claramente a cadeia causal em que o programa funciona, para gerar produtos e influenciar os resultados finalísticos, e que combinem avaliações de impacto com monitoramento e abordagens de avaliação complementares para obterem uma visão completa do desempenho do programa.

A maior novidade deste livro talvez seja a abordagem aplicada das ferramentas de avaliação de impacto ao trabalho de desenvolvimento que se dá no mundo real. As nossas experiências e lições sobre como fazer avaliação de impacto, na prática, advêm da atividade de ensino e do trabalho com centenas de parceiros competentes em governos, universidades e agências de desenvolvimento. Considerando todos os seus autores, o livro é fruto de décadas de experiência de trabalho com avaliações de impacto em quase todos os cantos do mundo.

Este livro se vale de um conjunto básico de materiais de ensino desenvolvidos para as oficinas "Transformando Promessas em Evidências", organizadas pelo gabinete do Economista-Chefe para o Desenvolvimento Humano (HDNCE), em parceria com unidades regionais e com o Grupo de Pesquisa em Desenvolvimento Econômico (DECRG) do Banco Mundial. No momento da redação, a oficina já havia sido realizada mais de 20 vezes em todas as regiões do mundo. As oficinas e este manual só foram possíveis graças às generosas contribuições do Governo Espanhol e do Departamento para o Desenvolvimento Internacional do Reino Unido (DfID) e ao Fundo Estratégico de Avaliação de Impacto (SIEF). Este manual e as apresentações e palestras que o acompanham estão disponíveis no site http://www.worldbank.org/ieinpractice.

Outras fontes de alta qualidade que fornecem introduções à avaliação de impacto são, por exemplo, Baker 2000; Ravallion 2001, 2008, 2009; Duflo,

Glennerster e Kremer 2007; Duflo e Kremer 2008; Khandker, Kool-wal e Samad 2009; e Leeuw e Vaessen 2009. O presente livro se diferencia por combinar uma visão geral abrangente e não técnica dos métodos quantitativos de avaliação de impacto com uma ligação direta com as regras de operação dos programas, assim como traz uma discussão detalhadas dos aspectos práticos de implementação. O livro também está vinculado a um curso sobre avaliação de impacto e a material de apoio para capacitação.

Os materiais de ensino nos quais o livro se baseia têm passado por muitas gerações e têm sido usados em sala de aula por vários acadêmicos talentosos que imprimiram sua marca nos métodos e abordagens de avaliação de impacto. Paul Gertler e Sebastian Martinez, juntamente com Sebastian Galiani e Sigrid Vivo, reuniram um primeiro conjunto de materiais de ensino para uma oficina ministrada no Ministério do Desenvolvimento Social (SEDESOL) no México, em 2005. Christel Vermeersch desenvolveu e refinou extensas sessões dos módulos técnicos da oficina e adaptou um estudo de caso para o contexto da oficina. Laura Rawlings e Patrick Premand desenvolveram materiais usados nas versões mais recentes da oficina.

Gostaríamos de agradecer e reconhecer as contribuições e insumos substanciais de uma série de outros acadêmicos que têm ministrado a oficina, incluindo Felipe Barrera, Sergio Bautista-Arredondo, Stefano Bertozzi, Barbara Bruns, Pedro Carneiro, Nancy Qian, Jishnu Das, Damien de Walque, David Evans, Claudio Ferraz, Jed Friedman, Emanuela Galasso, Sebastian Galiani, Gonzalo Hernández Licona, Arianna Legovini, Phillippe Leite, Mattias Lundberg, Karen Macours, Plamen Nikolov, Berk Özler, Gloria M. Rubio e Norbert Schady. Agradecemos os comentários de nossos revisores, Barbara Bruns, Arianna Legovini, Dan Levy, e Emmanuel Skoufias, bem como de Bertha Briceno, Gloria M. Rubio e Jennifer Sturdy. Também reconhecemos e agradecemos os esforços da talentosa equipe organizadora das oficinas, formada por Paloma Acevedo, Theresa Adobea Bampoe, Febe Mackey, Silvia Paruzzolo, Tatyana Ringland, Adam Ross, Jennifer Sturdy, e Sigrid Vivo.

Os originais sobre os quais se baseiam partes deste livro foram escritos durante oficina ministrada em Pequim, China, em julho de 2009. Agradecemos a todos que participaram no esboço da transcrição original da oficina, em particular: Paloma Acevedo, Carlos Asenjo, Sebastian Bauhoff, Bradley Chen, Changcheng Song, Jane Zhang e Shufang Zhang. Agradecemos, também, a Kristine Cronin por sua excelente assistência na pesquisa, a Marco Guzman e Martin Ruegenberg, por elaborarem as ilustrações, e a Cindy A. Fisher, Fiona Mackintosh e Stuart K. Tucker, por seu apoio editorial durante a produção do livro.

Reconhecemos com gratidão o apoio de todo o Banco Mundial a esta linha de trabalho, incluindo o apoio e liderança de Ariel Fiszbein, Arianna Legovini e Martin Ravallion.

A versão em Português foi traduzida por Translanguage. Somos especialmente gratos a Andre Loureiro, Leandro Costa, Bénédicte de la Brière, Phillipe Leitte, Barbara Bruns, Armando Amorim Simões, Judas Baloi, Julieta Trias, Holly Blagrave, Aliza Marcus e Daphna Berman. O financiamento para a tradução foi provido pelo Fundo Estratégico de Avaliação de Impacto (SIEF), com generoso aporte do Departamento para o Desenvolvimento Internacional do Governo Britânico.

Finalmente, gostaríamos de agradecer aos participantes das oficinas realizadas na Cidade do México, Nova Delhi, Cuernavaca, Ankara, Buenos Aires, Paipa, Fortaleza, Sofia, Cairo, Managua, Madri, Washington, Manila, Pretoria, Tunis, Lima, Amman, Pequim, Sarajevo, Cidade do Cabo, San Salvador, Kathmandu, Rio de Janeiro e Acra. Através de seu interesse, perguntas rigorosas e entusiasmo, pudemos aprender, passo a passo, o que os gestores de políticas buscam nas avaliações de impacto. Esperamos que este livro reflita suas ideias.

Referências

Baker, J. (2000). *Evaluating the Impact of Development Projects on Poverty. A Handbook for Practitioners*. Washington, DC: Banco Mundial.

Duflo, E., Glennerster, R. & Kremer, M. (2007). Using Randomization in Development Economics Research: A Toolkit. *CEPR Discussion Paper Nº 6059*. Reino Unido: Centro para Pesquisa de Política Econômica.

Duflo, E. & Kremer, M. (2008). Use of Randomization in the Evaluation of Development Effectiveness. Em *Evaluating Development Effectiveness*, vol. 7 Washington, DC: Banco Mundial.

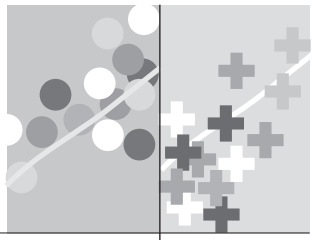
Khandker, S. R., Koolwal, G. & Samad, H. (2009). *Handbook on Quantitative Methods of Program Evaluation*. Washington, DC: Banco Mundial.

Leeuw, F. & Vaessen, J. (2009). *Impact Evaluations and Development. NONIE Guidance on Impact Evaluation*. Washington DC: NONIE e Banco Mundial.

Ravallion, Martin. (2001). The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation. *World Bank Economic Review* 15 (1): 115–40.

——. (2008). Evaluating Anti-Poverty Programs. Em *Manual de Desenvolvimento Econômico,* vol 4, ed. Schultz, P. & Strauss, J. Amsterdam.

——. (2009). Evaluation in the Practice of Development. *World Bank ResearchObserver* 24 (1): 29–53.



Parte 1

INTRODUÇÃO À AVALIAÇÃO DE IMPACTO

Nesta primeira parte do livro, damos uma visão geral sobre o que é avaliação de impacto. No capítulo 1, discutimos por que a avaliação de impacto é importante e como ela se encaixa no contexto da implementação de políticas com base em evidências. Contrastamos a avaliação de impacto com outras práticas comuns de avaliação, tais como monitoramento e avaliação de processo. Finalmente, apresentamos diferentes modalidades de avaliação de impacto, tais como a avaliação prospectiva e retrospectiva e testes de eficácia versus eficiência.

No capítulo 2, discutimos como formular perguntas de avaliação e hipóteses que sejam úteis às políticas. Estas questões e hipóteses formam a base da avaliação, pois determinam o que a avaliação busca responder.

CAPÍTULO 1

Por que Avaliar?

Os programas e políticas de desenvolvimento são, normalmente, elaborados para alterar resultados como, por exemplo, aumentar a renda, melhorar o aprendizado ou reduzir doenças. Se estas mudanças são, realmente, alcançadas ou não é uma questão crucial de política pública, mas que não é examinada com frequência. Gestores de programas e formuladores de políticas concentram-se, mais comumente, em controlar e medir os insumos e produtos imediatos de um programa, tais como o total de gastos ou o número de livros escolares distribuídos, ao invés de avaliar se os programas atingiram os objetivos pretendidos de melhoria do bem-estar.

Formulação de Políticas com Base em Evidências

As avaliações de impacto fazem parte de uma agenda mais ampla: a da formulação de políticas com base em evidências. Essa crescente tendência global está marcada por uma mudança no enfoque, de insumos para resultados. Dos Objetivos de Desenvolvimento do Milênio, passando pelos incentivos por desempenho que são pagos a prestadores de serviços públicos, essa tendência global está remodelando a forma de execução das políticas públicas. O foco nos resultados está sendo usado não somente para definir e monitorar metas nacionais e internacionais; de forma crescente, vem também sendo usado e exigido por gestores de programas para aperfeiçoar a

prestação de contas, informar a alocação orçamentária e orientar as decisões relacionadas a políticas.

O monitoramento e a avaliação estão no cerne da formulação de políticas com base em evidências. Fornecem um núcleo básico de ferramentas que as partes interessadas podem usar para verificar e melhorar a qualidade, eficiência e efetividade das intervenções nas várias etapas de execução ou, em outras palavras, focar em resultados. Podem ser encontrados agentes que fazem uso do monitoramento e da avaliação tanto em organismos governamentais quanto fora deles. Em agências governamentais ou ministérios, os agentes públicos geralmente precisam argumentar e defender o funcionamento de determinado programa junto a seus superiores e, assim, obter recursos para continuá-lo ou expandi-lo. Em nível nacional, os ministérios concorrem uns com os outros por recursos do Ministério da Fazenda. E, finalmente, os governos como um todo também têm interesse em convencer seus eleitores de que os investimentos realizados têm retornos positivos. Neste sentido, a informação e as evidências se tornam meios para facilitar a conscientização do público e promover a responsabilidade governamental. As informações produzidas pelos sistemas de monitoramento e avaliação podem ser divulgadas regularmente aos cidadãos, para informar-lhes sobre o desempenho de programas governamentais e construir uma base sólida de transparência e prestação de contas.

Em um contexto no qual os formuladores de políticas e a sociedade civil exigem resultados e a cobram prestação de contas dos programas públicos, a avaliação de impacto pode oferecer evidências críveis e robustas quanto ao desempenho e, fundamentalmente, quanto a se um programa específico atingiu os resultados desejados. Em nível mundial, as avaliações de impacto são centrais à construção do conhecimento sobre a efetividade de programas de desenvolvimento, esclarecendo o que funciona e o que não funciona na diminuição da pobreza e na promoção do bem-estar.

Simplificando, uma avaliação de impacto avalia as mudanças no bem-estar dos indivíduos que podem ser *atribuídas* a um projeto, programa ou política em particular. Este enfoque na atribuição do resultado é o selo distintivo das avaliações de impacto. Igualmente, o desafio central da execução de avaliações de impacto é identificar a *relação causal* entre o projeto, programa ou política e os resultados de interesse.

Conforme será discutido adiante, as avaliações de impacto geralmente estimam os impactos médios de um programa sobre o bem-estar dos beneficiados. Por exemplo, a introdução de um novo currículo aumentou a pontuação nos testes dos estudantes? Um programa de água e saneamento aumentou o acesso à água tratada e melhorou os resultados de saúde? Um programa de treinamento de jovens foi efetivo em estimular o

empreendedorismo e o aumento da renda? Além disso, se uma avaliação de impacto inclui uma amostra suficientemente grande de beneficiários, os resultados também podem ser comparados entre subgrupos de beneficiários. Por exemplo, a introdução de um novo currículo aumentou a pontuação nos testes entre estudantes do sexo masculino e estudantes do sexo feminino? As avaliações de impacto também podem ser usadas para testar explicitamente as opções de programas alternativos. Por exemplo, uma avaliação pode comparar o desempenho de um programa de treinamento versus o desempenho de uma campanha promocional para desenvolver a educação financeira. Em cada um destes casos, a avaliação de impacto fornece informações sobre o impacto geral de um programa, em contraste com estudos de casos específicos, que podem fornecer apenas informações parciais e podem não ser representativos dos impactos gerais de determinado programa. Neste sentido, avaliações bem formuladas e bem executadas podem fornecer evidências abrangentes e convincentes, que podem ser usadas para informar as decisões sobre políticas e formar a opinião pública. O resumo no Quadro 1.1 ilustra como a avaliação de impacto contribuiu para o debate sobre a expansão de um programa de transferência condicionada de renda no México[1]. O Quadro 1.2 ilustra como a avaliação de impacto ajudou a melhorar a alocação de recursos do governo da Indonésia, documentando quais políticas foram mais efetivas na diminuição das taxas de fertilidade.

O que é Avaliação de Impacto?

A avaliação de impacto figura entre uma série abrangente de métodos complementares que oferecem suporte às políticas baseadas em evidências. Embora este livro enfoque nos métodos quantitativos de avaliação de impacto, começaremos colocando estes métodos no contexto mais amplo de resultados, o que também inclui monitoramento e outros tipos de avaliação.

O *monitoramento* é um processo contínuo, que acompanha o que está acontecendo com o programa e usa os dados coletados para informar a implementação do programa e a gestão e decisões tomadas no dia a dia. Utilizando, principalmente, dados administrativos, o monitoramento rastreia o desempenho do programa e o compara ao resultado esperado, realiza comparações entre programas e analisa tendências ao longo do tempo. Geralmente, o monitoramento cobre insumos, atividades e produtos, embora possa, às vezes, incluir resultados, tais como o progresso em direção às metas nacionais de desenvolvimento.

As *Avaliações* são processos periódicos, julgamentos objetivos de um projeto, programa ou política planejada, em andamento ou concluída.

Quadro 1.1: Avaliações e Sustentabilidade Política

O programa de transferência condicionada de renda *Progresa/Oportunidades* no México

Nos anos 90, o governo do México lançou um programa inovador de transferência condicionada de renda chamado "Progresa". Seu objetivo era dar às famílias pobres um apoio financeiro de curto prazo e criar incentivos para investimentos no capital humano das crianças, principalmente por meio da transferência de renda às mães de famílias pobres com a condição de que os seus filhos estejam frequentando regularmente a escola e comparecendo a um centro de saúde.

Desde o começo, o governo considerou que era fundamental monitorar e avaliar o programa. As autoridades do programa contrataram um grupo de pesquisadores para desenhar uma avaliação de impacto e integrá-la ao processo de expansão do programa, à medida que este fosse estendido às comunidades participantes.

As eleições presidenciais de 2000 trouxeram uma mudança de partido no poder. Em 2001, os avaliadores externos do "Progresa" apresentaram os resultados de seu estudo ao governo recém-eleito. Os resultados do programa eram impressionantes: eles mostraram que o programa era bem focalizado nos pobres e que gerou mudanças promissoras no capital humano das famílias. Schultz (2004) concluiu que o programa aumentou significativamente os anos de matrícula escolar, em uma média de 0,7 anos adicionais de estudo. Gertler (2004) constatou que a incidência de doenças infantis diminuiu 23%, enquanto os adultos relataram uma redução de 19% na quantidade de dias de doença ou incapacidade. Entre os resultados nutricionais, Behrman e Hoddinott (2001) relatam que o programa reduziu a probabilidade de baixo crescimento em cerca de 1 centímetro por ano para crianças na faixa etária crítica, de 12 a 36 meses.

Esses resultados da avaliação deram suporte a um diálogo político baseado em evidências e contribuíram para a decisão do novo governo de continuar com o programa. Por exemplo, o governo expandiu o alcance do programa, introduzindo bolsas para os alunos de nível médio e melhorando os programas de saúde para adolescentes. Ao mesmo tempo, os resultados foram usados para modificar outros programas assistenciais - como o subsídio *tortilla,* amplo e menos focalizado, que foi reduzido.

A avaliação bem-sucedida do Progresa contribuiu para a rápida adoção dos programas de transferência condicionada de renda em todo o mundo, assim como a adoção de legislação, no México, exigindo que todos os projetos sociais sejam avaliados.

Fontes: Behrman e Hoddinott 2001; Gertler 2004; Fiszbein e Schady 2009; Levy e Rodriguez 2005; Schultz 2004; Skoufias e McClafferty 2001.

Quadro 1.2: Avaliando para Melhorar a Alocação de Recursos
Planejamento familiar e fertilidade na Indonésia

Na década de 70, os esforços inovadores do planejamento familiar na Indonésia ganharam reconhecimento internacional por seu sucesso em diminuir as taxas de fertilidade do país. A notoriedade surgiu de dois fenômenos paralelos: (1) as taxas de fertilidade apresentaram uma queda, de 22% entre os anos de 1970 e 1980, de 25% entre os anos de 1981 e 1990 e um pouco mais moderada entre os anos de 1991 e 1994; e (2) durante o mesmo período, o governo da Indonésia aumentou consideravelmente os recursos alocados ao planejamento familiar (particularmente, os subsídios para anticoncepcionais). Dado que as duas coisas ocorreram simultaneamente, muitos concluíram que foi o aumento do investimento no planejamento familiar que levou a um índice de fertilidade mais baixo.

Não convencidos pela evidência disponível, uma equipe de pesquisadores testou se os programas de planejamento familiar realmente diminuíram as taxas de fertilidade. Eles descobriram, ao contrário do que se acreditava, que os programas de planejamento familiar tiveram um impacto apenas moderado sobre a fertilidade e argumentaram que, ao invés disso, foi a mudança no status das mulheres que teria sido responsável pelo declínio na taxa de fertilidade. Os pesquisadores notaram que, antes do início do programa de planejamento familiar, muito poucas mulheres na idade fértil haviam concluído o ensino fundamental. Durante o mesmo período do programa de planejamento familiar, no entanto, o governo executou um amplo programa de educação para meninas, de modo que, ao final do programa, as mulheres que entravam na idade fértil haviam se beneficiado de mais anos de educação. Quando o crescimento do setor petrolífero expandiu a economia e aumentou a demanda por trabalho na Indonésia, a participação de mulheres escolarizadas no mercado de trabalho cresceu consideravelmente. Conforme aumentou o valor do tempo de trabalho da mulher, aumentou também o uso de anticoncepcionais. Ao final, salários mais altos e empoderamento das mulheres foram responsáveis por 70% do declínio observado na fertilidade, mais do que o investimento em programas de planejamento familiar.

Estes resultados de avaliação informaram as decisões sobre a alocação de recursos tomadas posteriormente pelos gestores de políticas: o financiamento foi reprogramado, dos subsídios aos anticoncepcionais para os programas de estímulo à matrícula escolar das mulheres. Embora os objetivos finais dos dois tipos de programa fossem semelhantes, os estudos de avaliação demonstraram que, no contexto da Indonésia, taxas mais baixas de fertilidade poderiam ser alcançadas de modo mais eficiente através de investimentos em educação, ao invés de investimentos em planejamento familiar.

Fontes: Gertler e Molyneaux 1994, 2000.

As avaliações são usadas para responder a perguntas específicas relacionadas à formulação, implementação e obtenção de resultados. Ao contrário do monitoramento contínuo, elas são realizadas em pontos discretos no tempo e, geralmente, buscam uma perspectiva externa de especialistas. O seu desenho, método e custo variam substancialmente, dependendo do tipo de pergunta que a avaliação tenta responder. Em geral, as avaliações podem abordar três tipos de questões (Imas e Rist 2009):

- *Perguntas descritivas*. A avaliação procura determinar o que está sendo executado e descreve processos, condições, relações organizacionais e pontos de vista das partes interessadas.
- *Perguntas normativas*. A avaliação compara o que está sendo executado ao que deveria ocorrer; avalia atividades e se as metas estão ou não sendo alcançadas. As perguntas normativas podem ser aplicadas a insumos, atividades ou produtos.
- *Perguntas de causa e efeito*. A avaliação analisa resultados e procura identificar que diferença a intervenção faz para os resultados.

As avaliações de impacto são um tipo particular de avaliação, que procura responder a perguntas de causa e efeito. Diferentemente das avaliações gerais, que podem responder a muitos tipos de perguntas, as avaliações de impacto se estruturam em torno de um tipo específico de pergunta: *qual é o impacto (ou efeito causal) de um programa sobre um resultado de interesse?* Esta pergunta básica incorpora uma importante dimensão causal: estamos interessados apenas no *impacto* do programa, isto é, no efeito que o programa causa diretamente no resultado. Uma avaliação de impacto busca as mudanças nos resultados que são *atribuíveis diretamente ao programa*.

O enfoque na causalidade e atribuição de resultados é a característica distintiva das avaliações de impacto e determina as metodologias que podem ser usadas. Para poder estimar o efeito causal ou impacto de um programa sobre os resultados, qualquer método escolhido deve estimar o chamado contrafactual, isto é, qual teria sido o resultado para os participantes do programa se eles não tivessem participado do programa. Na prática, a avaliação de impacto exige que o avaliador encontre um grupo de comparação para estimar o que teria acontecido aos participantes do programa na ausência do programa. A segunda parte do livro descreve os principais métodos que podem ser usados para encontrar grupos de comparação adequados.

Conceito-chave:
A pergunta básica da avaliação de impacto pode ser formulada como: qual é o impacto (ou efeito causal) de um programa sobre um resultado de interesse?

A pergunta básica da avaliação de impacto, *"qual é o impacto (ou efeito causal) de um programa sobre um resultado de interesse?"* pode ser aplicada a muitos contextos. Por exemplo, qual é o efeito causal das bolsas na frequência escolar e desempenho acadêmico dos alunos? Qual é o impacto no acesso

aos serviços de saúde da contratação de serviços de atenção básica de fornecedores privados? Se o chão de terra batida fosse substituído por chão de cimento, qual seria o impacto na saúde das crianças? Será que estradas melhores aumentam o acesso aos mercados de trabalho e a renda familiar? Caso aumentem, em quanto? O tamanho das turmas influencia o desempenho dos alunos? Caso influencie, em quanto? O que é mais efetivo para o aumento do uso de mosquiteiros em áreas afetadas pela malária, campanhas pelo correio ou sessões de treinamento?

Avaliação de Impacto em Decisões de Políticas

As avaliações de impacto são necessárias para informar os formuladores de políticas em várias decisões, desde a redução de programas ineficientes até a expansão de intervenções que funcionam, a adequação de benefícios do programa e a seleção entre várias alternativas de programas. Elas são mais efetivas quando aplicadas seletivamente para responder a perguntas importantes de políticas, e podem ser particularmente efetivas quando aplicadas a programas piloto inovadores que avaliam uma nova abordagem ainda não testada, porém promissora. A avaliação do programa mexicano Progresa/Oportunidades, descrita no Quadro 1.1, se tornou bastante influente, não somente devido à natureza inovadora do programa, mas também porque sua avaliação de impacto trouxe evidências fortes e convincentes, que não poderiam ser ignoradas em decisões de política subsequentes. A adoção e a extensão do programa foram fortemente influenciadas pelos resultados da avaliação. Hoje, o programa Oportunidades atinge cerca de 1 em cada 4 mexicanos e é uma peça central da estratégia do México para combater a pobreza.

As avaliações de impacto podem ser usadas para explorar diferentes tipos de questões sobre políticas. A forma básica de avaliação de impacto testará a efetividade de um determinado programa. Em outras palavras, responderá à pergunta: *determinado programa é efetivo quando comparado à sua própria ausência?* Conforme apresentado na segunda parte, este tipo de avaliação de impacto depende da comparação de um grupo de tratamento que recebeu o projeto, programa ou política com um grupo de comparação que não o recebeu, a fim de estimar a efetividade do programa.

Além de responder à pergunta básica da avaliação, as avaliações também podem ser usadas para testar a efetividade das alternativas de implementação do programa - isto é, para responder à pergunta: *quando um programa puder ser implementado de várias formas, qual delas será a mais efetiva?* Neste tipo de avaliação, duas ou mais abordagens dentro de um mesmo programa podem ser comparadas para gerar evidências sobre qual é a melhor

alternativa para se atingir um objetivo específico. Estas alternativas de programa são, geralmente, chamadas de "braços de tratamento". Por exemplo, quando a quantidade de benefícios que um programa deve fornecer não estiver clara (20 ou 80 horas de capacitação?), as avaliações de impacto podem testar o impacto relativo de intensidades variadas de tratamento (veja um exemplo no Quadro 1.3). Avaliações de impacto que testam tratamentos alternativos de programas geralmente incluem um grupo de tratamento para cada um dos tipos (ou "braços") de tratamento, bem como um

Quadro 1.3: Avaliando para Melhorar o Desenho do Programa

Desnutrição e desenvolvimento cognitivo na Colômbia

No início dos anos 70, a Estação de Pesquisa de Ecologia Humana implementou, em colaboração com o Ministério da Educação colombiano, um programa piloto para enfrentar a situação da desnutrição infantil em Cali, na Colômbia, provendo cuidados à saúde e atividades educativas, assim como suplementos nutricionais e alimentícios. Como parte do piloto, uma equipe de avaliadores foi encarregada de determinar (1) quanto tempo o programa deveria durar para reduzir a desnutrição entre crianças de idade pré-escolar em famílias de baixa renda e (2) se as intervenções também poderiam resultar em melhorias no desenvolvimento cognitivo.

O programa foi disponibilizado para todas as famílias elegíveis e, durante o piloto, os avaliadores puderam comparar grupos semelhantes de crianças que receberam diferentes durações do tratamento. Os avaliadores usaram, primeiramente, um processo de filtragem, para identificar um grupo-alvo de 333 crianças desnutridas. Estas crianças foram, então, classificadas por vizinhança em 20 setores; cada setor foi aleatoriamente designado a um dos quatro grupos de tratamento. Os grupos se diferenciavam somente na sequência em que começavam o tratamento e, portanto, na quantidade de tempo que permaneceram no programa. O grupo 4 começou antes e foi exposto ao tratamento pelo período mais longo, seguido pelos grupos 3, 2 e, por último, 1. O tratamento em si consistiu em 6 horas diárias de cuidados com a saúde e atividades educativas, além de alimentos adicionais e suplementos nutricionais. Em intervalos regulares durante o programa, os avaliadores usaram testes cognitivos para acompanhar o progresso das crianças em todos os quatro grupos.

Os avaliadores descobriram que as crianças que permaneceram no programa por mais tempo demonstraram os maiores ganhos em melhoria cognitiva. No teste de inteligência de Stanford-

Binet, que estima a idade mental menos a idade cronológica, o grupo 4 obteve uma média de -5 meses e o grupo 1 obteve uma média de -15 meses.

Este exemplo ilustra como os implementadores de programas e formuladores de políticas podem usar as avaliações de múltiplos tratamentos para determinar a alternativa mais efetiva de programa.

Fonte: McKay et al. 1978.

grupo de comparação "limpo" e que não recebe nenhuma intervenção do programa. As avaliações de impacto também podem ser usadas para testar inovações ou alternativas de implementação de um determinado programa. Por exemplo, um programa pode desejar testar alternativas de campanhas de sensibilização - selecionando um grupo para receber a campanha pelo correio enquanto os outros recebem visitas de casa em casa - para avaliar qual é a mais efetiva.

Decidindo se Deve ou não Avaliar

Nem todos os programas justificam uma avaliação de impacto. As avaliações de impacto podem ser onerosas e o orçamento para tal deve ser usado estrategicamente. Se você estiver iniciando ou considerando expandir um novo programa e estiver refletindo se é o momento de ir em frente com uma avaliação de impacto, algumas perguntas básicas ajudarão na decisão.

A primeira pergunta a ser feita é: *O que está em questão no programa?* A resposta a esta pergunta dependerá tanto do orçamento como da quantidade de pessoas que são ou serão afetadas pelo programa. Consequentemente, as próximas questões são: *o programa requer ou requererá uma grande parte do orçamento disponível? O programa afeta ou afetará um grande número de pessoas?* Se o programa não requer orçamento ou afetar somente algumas pessoas, pode ser que não valha a pena avaliar. Por exemplo, no caso de um programa que fornece aconselhamento a pacientes em hospitais por meio de voluntários, o orçamento envolvido e a quantidade de pessoas afetadas pode não justificar uma avaliação de impacto. Por outro lado, uma reforma salarial para o magistério, que eventualmente venha a impactar todos os professores do Ensino Fundamental de um determinado país, caracteriza um programa com riscos muito mais altos.

Se você determinar que o programa é suficientemente significativo então, a próxima pergunta é se há alguma evidência que mostre que o programa funciona. Em particular, você sabe qual seria o tamanho do impacto do programa? Há evidências disponíveis de um país semelhante com circunstâncias semelhantes? Se não houver nenhuma evidência disponível a respeito do potencial do programa sob consideração, você pode querer iniciar um piloto que incorpore uma avaliação de impacto. Por outro lado, se existirem evidências disponíveis envolvendo circunstâncias semelhantes, o custo de uma avaliação de impacto provavelmente só se justificará se puder abordar uma questão de política nova e relevante. Este seria o caso se o seu programa incluísse inovações importantes e que ainda não puderam ser testadas.

Para justificar a mobilização dos recursos técnicos e financeiros necessários para executar uma avaliação de impacto de alta qualidade, o programa a ser avaliado deve ser:

- *Inovador*: está testando uma abordagem nova e promissora.
- *Replicável*: o programa pode ser expandido ou pode ser aplicado em um contexto diferente.
- *Estrategicamente relevante*: o programa é uma iniciativa emblemática; requer recursos consideráveis; cobre - ou poderia ser expandido para cobrir - um grande número de pessoas; ou poderia gerar grandes economias.
- *Ainda não testado*: pouco se sabe sobre a efetividade do programa, em contexto global ou particular.
- *Influente*: os resultados serão usados para informar decisões cruciais sobre políticas.

Análise de Custo-Efetividade

Uma vez que os resultados da avaliação de impacto estejam disponíveis, eles podem ser combinados com informações sobre os custos do programa, para responder a duas perguntas adicionais. Primeiramente, considerando a forma básica da avaliação de impacto, adicionar informações de custo nos permitirá realizar uma análise de custo-benefício, que responderá à pergunta: *qual é a relação custo-benefício de um determinado programa?* A análise custo-benefício estima o total de benefícios esperados de um programa, comparado ao total de custos esperados. Ela procura quantificar, em termos monetários, todos os custos e benefícios de um programa e avaliar se os benefícios ultrapassaram os custos.

Conceito-chave:
A análise custo-benefício estima o total de benefícios esperados de um programa, comparado ao total de custos esperados.

Em um mundo ideal, a análise de custo-benefício baseada em evidências de avaliação de impacto ocorreria não somente para um programa específico, mas também para uma série de programas ou alternativas de programa, de modo que os formuladores de políticas pudessem avaliar qual programa ou alternativa de programa teria a melhor relação custo-efetividade para alcançar um determinado objetivo. Quando uma avaliação de impacto testa alternativas de programa, a adição de informações de custo nos permite responder uma segunda pergunta: *como as várias alternativas de implementação do programa se comparam, em termos de custo-efetividade?* Esta análise de custo-efetividade compara o desempenho relativo de dois ou mais programas (ou alternativas de programa) para alcançar um mesmo resultado.

Conceito-chave:
A análise de custo-efetividade compara o desempenho relativo de dois ou mais programas - ou de alternativas de um programa para alcançar um mesmo resultado.

Em uma análise de custo-benefício ou custo-efetividade, a avaliação de impacto estima o lado do benefício e da efetividade e a análise de custos fornece informações sobre o custo. Este livro enfoca a avaliação de impacto e não discute detalhadamente como coletar dados de custos ou realizar a análise de custo-benefício[2]. Entretanto, é fundamental que a avaliação de impacto seja complementada por informações sobre os custos do projeto, do programa ou da política que estiver sendo avaliada. Uma vez que as informações de impactos e custos estejam disponíveis para uma variedade de programas, a análise de custo-efetividade pode identificar quais investimentos geram as taxas mais altas de retorno e permite aos gestores de políticas tomarem decisões informadas sobre as intervenções em que devem investir. O Quadro 1.4 ilustra como as avaliações de impacto podem ser usadas para identificar os programas de melhor custo-efetividade e melhorar a alocação de recursos.

Quadro 1.4: Avaliando a Custo-Efetividade

Comparando estratégias para aumentar a frequência escolar no Quênia

Ao avaliar uma variedade de programas em um contexto semelhante, é possível comparar o custo-efetividade relativo de alternativas que buscam melhorar resultados tais como a frequência escolar. No Quênia, a organização não governamental *International Child Support Africa* (ICS Africa) implementou uma série de intervenções educacionais que incluíam tratamento contra parasitas intestinais e uniformes e merenda escolar gratuitos. Cada uma das intervenções passou por uma avaliação de natureza aleatória e uma análise de custo benefício - a comparação entre elas traz conhecimentos interessantes sobre como aumentar a frequência escolar.

O programa de medicamentos contra parasitas intestinais para crianças em idade escolar aumentou a frequência em aproximadamente 0,14 anos por criança tratada, a um custo estimado de $0,49 por criança. Isto equivale a aproximadamente $3,50 por ano adicional de participação escolar, incluindo as externalidades experimentadas por crianças e adultos que, embora não estejam na escola, vivem nas comunidades beneficiadas pela redução na transmissão de parasitas.

Uma segunda intervenção, o programa de patrocínio infantil, reduziu o custo da frequência escolar ao fornecer uniformes escolares a alunos em sete escolas selecionadas aleatoriamente. As taxas de evasão caíram dramaticamente nas escolas tratadas e, após 5 anos, estima-se que o programa aumente em 17%, em média, os anos de matrícula escolar. No entanto, mesmo sob as hipóteses mais otimistas, o custo de aumentar a frequência escolar usando o programa de uniforme escolar foi estimado em aproximadamente $99 por ano adicional de frequência escolar.

Finalmente, o programa de café da manhã gratuito oferecido a crianças em 25 pré-escolas

(continua)

Quadro 1.3 *continuação*

selecionadas aleatoriamente levou a um aumento de 30% na frequência nas escolas de tratamento, a um custo estimado de $36 por ano adicional de escolarização. Os resultados dos testes de aprendizagem também aumentaram em 0,4 desvios-padrão, contanto que os professores tenham sido bem treinados antes do programa.

Embora intervenções semelhantes possam ter objetivos diferentes - tais como os efeitos do tratamento das verminoses na saúde ou desempenho escolar devido a uma maior frequência escolar - comparar uma série de avaliações realizadas no mesmo contexto pode revelar quais programas alcaçaram o objetivo desejado, com os menores custos.

Fontes: Kremer e Miguel, 2004; Kremer, Moulin e Namunyu, 2003; Poverty Action Lab, 2005; Vermeersch e Kremer, 2005.

Avaliação Prospectiva versus Avaliação Retrospectiva

Conceito-chave:
Avaliações prospectivas são desenvolvidas ao mesmo tempo em que o programa está sendo elaborado e são integradas à implementação do programa

As avaliações de impacto podem ser divididas em duas categorias: prospectivas e retrospectivas. Avaliações prospectivas são desenvolvidas ao mesmo tempo em que o programa está sendo elaborado e são integradas à implementação do programa. Dados de linha de base são coletados antes da implementação do programa, tanto para grupos de tratamento quando de comparação. As avaliações retrospectivas avaliam o impacto do programa após sua implementação, gerando grupos de tratamento e de comparação *a posteriori*.

Em geral, as avaliações de impacto prospectivas têm maior probabilidade de produzir resultados mais robustos e confiáveis, por três razões.

Em primeiro lugar, podem ser coletados dados de linha de base para estabelecer as medidas pré-programa dos resultados de interesse. Os dados de linha de base fornecem informações sobre beneficiários e grupos de comparação antes da implementação do programa e são importantes para medir os resultados pré-intervenção. Devem ser analisados os dados de linha de base dos grupos de tratamento e de comparação para garantir que os grupos sejam semelhantes. As linhas de base também podem ser usadas para avaliar a efetividade da focalização - isto é, se o programa vai, ou não, atingir os beneficiários pretendidos.

Em segundo lugar, definir medidas de sucesso de um programa na etapa de planejamento ajuda a focar a avaliação e o próprio programa nos resultados pretendidos. Como veremos, as avaliações de impacto têm origem na teoria de mudança do programa ou cadeia de resultados. O desenho de uma avaliação de impacto ajuda a esclarecer os objetivos do programa,

em particular porque requer o estabelecimento de medidas bem definidas do sucesso do programa. Os formuladores de políticas devem estabelecer objetivos claros e perguntas de avaliação para garantir que os resultados sejam altamente relevantes em termos de política. Na verdade, o apoio integral dos formuladores de políticas é um pré-requisito para executar uma avaliação bem-sucedida; as avaliações de impacto não devem ser realizadas a menos que os formuladores de políticas estejam convencidos da legitimidade da avaliação e de seu valor, no sentido de informar decisões sobre políticas.

Em terceiro lugar, o mais importante: em uma avaliação prospectiva, os grupos de tratamento e de comparação são identificados antes que o programa seja implementado. Como explicaremos mais detalhadamente nos capítulos seguintes, existem muito mais opções para executar avaliações válidas quando as avaliações são planejadas desde o princípio e integradas à implementação do projeto. Nas partes 2 e 3, discutimos que quase sempre pode ser encontrada uma estimativa válida do cenário contrafactual para qualquer programa que tenha regras de atribuição claras e transparentes, desde que a avaliação seja elaborada prospectivamente. Em suma, as avaliações prospectivas têm maiores chances de gerar contrafatuais válidos. Na fase de desenho, podem ser consideradas formas alternativas de estimar-se um contrafactual válido. O desenho da avaliação de impacto também pode ser integralmente alinhado às regras de operação do programa, assim como às etapas de implementação e expansão do programa.

Por outro lado, nas avaliações retrospectivas, o avaliador geralmente possui informações tão limitadas que se torna difícil analisar se o programa foi implementado com sucesso ou se os seus participantes realmente extraíram benefícios dele. Em parte, a razão disto se deve ao fato de que muitos programas não coletam dados de linha de base a menos que a avaliação tenha sido integrada desde o início e, uma vez que o programa começa a funcionar, é tarde demais para isso.

As avaliações retrospectivas que usam dados existentes são necessárias para avaliar os programas que foram alocados no passado. De modo geral, as opções para se obter uma estimativa válida do cenário contrafactual são muito mais limitadas nessas situações. A avaliação depende de regras claras de operação do programa referentes à alocação dos benefícios. Também depende da disponibilidade de dados, com cobertura suficiente dos grupos de tratamento e de comparação, tanto antes como depois da implementação do programa. Como resultado, a possibilidade de uma avaliação retrospectiva depende do contexto e nunca é garantida. Mesmo quando são possíveis de realizar, as avaliações retrospectivas geralmente lançam mão de métodos quase-experimentais, que dependem de fortes pressupostos; assim, elas tendem a produzir evidências mais facilmente contestáveis.

Estudos de Eficácia e de Efetividade

O principal papel da avaliação de impacto é produzir evidências sobre a efetividade de um programa para o uso de autoridades governamentais, gestores de programas, sociedade civil e outras partes interessadas. Os resultados da avaliação de impacto são particularmente úteis quando as conclusões podem ser aplicadas de forma mais ampla à população de interesse. A questão de generalização (conhecida como "validade externa" na literatura de métodos de pesquisa) é essencial para os tomadores de decisão, pois determina se os resultados identificados na avaliação podem ser replicados para outros grupos além daqueles estudados na avaliação, caso o programa venha a ser ampliado.

Nos primórdios das avaliações de impacto de programas de desenvolvimento, uma grande parcela de evidências se baseava em *estudos de eficácia* realizados sob circunstâncias muito específicas; infelizmente, os resultados daqueles estudos não podiam, com frequência, ser generalizados para além do escopo da avaliação. Os estudos de eficácia são, geralmente, realizados em um cenário muito específico, com forte envolvimento técnico de pesquisadores durante a execução do programa. Estes estudos de eficácia geralmente são realizados para provar conceitos e testar a viabilidade de um novo programa. Se o programa não gera os impactos antecipados sob tais condições, geralmente administradas com muito cuidado, é improvável que funcione sob circunstâncias normais. Por serem geralmente executados sob condições rigorosamente administradas, os impactos de tais pilotos de pequena escala podem não ser necessariamente informativos sobre o impacto de um projeto similar executado em uma escala maior e sob circunstâncias normais. Por exemplo, uma intervenção piloto que introduza novos protocolos de tratamento pode funcionar em um hospital com administradores e equipe médica excelentes, mas a mesma intervenção pode não funcionar em um hospital mediano, com administradores menos atentos e equipe limitada. Além disso, os cálculos de custo-benefício irão variar, considerando que os custos fixos e as economias de escala podem não ser captados em estudos de eficácia de pequena escala. Consequentemente, embora as evidências dos estudos de eficácia possam ser úteis para testar novas abordagens, os resultados geralmente apresentam validade externa limitada e nem sempre representam adequadamente os contextos mais gerais que, normalmente, configuram a principal preocupação dos gestores de políticas.

Por outro lado, *os estudos de efetividade* fornecem evidências a partir de intervenções que ocorrem em circunstâncias normais, pelo uso de canais regulares de execução. Quando as avaliações de efetividade são desenhadas e executadas adequadamente, os resultados obtidos se confirmarão não

somente para a amostra da avaliação, mas também para outros potenciais beneficiários fora da amostra. A validade externa é de fundamental importância para os formuladores de políticas, porque lhes permite usar os resultados da avaliação para informar decisões sobre o programa aplicáveis aos beneficiários além da amostra da avaliação.

Combinando Fontes de Informação para Avaliar "o Quê" e "Por Quê"

As avaliações de impacto realizadas isoladamente de outras fontes de informações são vulneráveis, tanto tecnicamente quanto em termos da sua efetividade potencial. Sem informações sobre a natureza e o conteúdo do programa para contextualizar os resultados da avaliação, os gestores de políticas ficam em dúvida em relação ao porquê de determinados resultados terem, ou não, sido atingidos. Enquanto as avaliações de impacto produzem estimativas confiáveis dos efeitos causais de um programa, elas não são elaboradas especificamente para gerar conhecimentos sobre a implementação do programa. Além disso, as avaliações de impacto devem ser bem alinhadas à implementação do programa e, consequentemente, precisam ser orientadas por informações sobre como, quando e onde o programa avaliado está sendo implementado.

Dados qualitativos, de monitoramento e avaliações de processos são necessários para rastrear a implementação do programa e examinar questões de processo que sejam críticas para informar e interpretar os resultados das avaliações de impacto. Neste sentido, as avaliações de impacto e outras formas de avaliação são complementares entre si, não substitutas.

Por exemplo, um governo provincial pode decidir anunciar que pagará bônus às clínicas médicas da zona rural se elas aumentarem a porcentagem de partos assistidos por um profissional da saúde na clínica. Se a avaliação concluir que não ocorreu mudança alguma na porcentagem de partos assistidos na clínica, podem existir muitas explicações possíveis e a cada uma corresponde uma ação necessária. Primeiro, pode ser que as equipes das clínicas rurais não possuam informações suficientes sobre os bônus ou não entendam as regras do programa. Neste caso, o governo provincial precisaria reforçar as informações sobre o bônus, além de oferecer campanhas educativas aos centros de saúde. Alternativamente, se a falta de equipamentos ou corte de energia impedir que as clínicas médicas recebam mais pacientes, pode ser necessário melhorar o sistema de suporte e melhorar o fornecimento de energia. Finalmente, as gestantes da zona rural podem não querer usar as clínicas; pode ser que elas prefiram parteiras e o parto feito em casa, por

razões culturais. Neste caso, pode ser mais eficaz derrubar as barreiras que as mulheres enfrentam em acessar as clínicas do que dar bônus às clínicas. Portanto, uma boa avaliação de impacto permitirá ao governo determinar se a taxa de partos assistidos mudou como resultado do programa de bônus. No entanto, são necessárias abordagens de avaliação complementares para entender se o programa foi implementado conforme o planejamento e onde faltam conexões. Neste exemplo, os avaliadores desejariam complementar sua análise de impacto entrevistando o pessoal da clínica de saúde em relação a seu conhecimento do programa, revisar a disponibilidade de equipamento nas clínicas, realizar grupos focais com gestantes para entender suas preferências e barreiras ao acessar o serviço clínico e examinar todos os dados disponíveis sobre o acesso aos centros de saúde na zona rural.

Usando Dados Qualitativos

Dados qualitativos são um suplemento-chave para as avaliações de impacto quantitativas, pois podem fornecer perspectivas complementares sobre o desempenho do programa. As avaliações que integram análises qualitativas e quantitativas são caracterizadas pelo uso de "métodos combinados" (Bamberger, Raim e Woolcock 2010). Abordagens qualitativas incluem grupos focais e entrevistas com beneficiários selecionados ou outros informantes-chave (Rai e Woolcock 2003). Embora os pontos de vista e opiniões reunidos durante as entrevistas e grupos focais possam não ser representativos dos beneficiários do programa, eles são particularmente úteis durante as três etapas de uma avaliação de impacto:

1. Durante o desenho da avaliação de impacto, os avaliadores podem usar grupos focais e entrevistas com informantes-chave para desenvolver hipóteses sobre como e por que o programa funcionaria e esclarecer questões de pesquisa que precisem ser abordadas no trabalho de avaliação quantitativa de impacto.

2. Na etapa intermediária, antes que os resultados da avaliação quantitativa de impacto estejam disponíveis, o trabalho qualitativo pode ajudar a fornecer aos gestores de políticas panoramas rápidos do que vem acontecendo no programa.

3. Na etapa de análise, os avaliadores podem aplicar métodos qualitativos para contextualizar e explicar os resultados quantitativos, para explorar casos "aberrantes" de sucesso e fracasso e desenvolver explicações sistemáticas do desempenho do programa, conforme expresso pelos resultados quantitativos. Neste sentido, o trabalho qualitativo pode

ajudar a explicar por que são observados determinados resultados na análise quantitativa e pode ser usado para examinar a "caixa preta" do que acontece no programa (Bamberger, Rao e Woolcock 2010).

Usando Dados de Monitoramento e Avaliações de Processo

Os dados de monitoramento também são um recurso crítico em uma avaliação de impacto. Eles permitem ao avaliador verificar quais participantes receberam o programa, com que rapidez o programa está se expandindo, como os recursos estão sendo gastos e, de maneira geral, se as atividades estão sendo implementadas conforme planejado. Estas informações são fundamentais ao implementar a avaliação, por exemplo, para garantir que os dados de linha de base sejam coletados antes que o programa seja introduzido e para verificar a integridade dos grupos de tratamento e de comparação. Além disso, o sistema de monitoramento pode fornecer informações sobre o custo de implementação do programa, que também são necessárias para a análise de custo-benefício.

Finalmente, *as avaliações de processo* enfocam como um programa é implementado e operado, avaliando se ele está em conformidade com o desenho original, e documenta seu desenvolvimento e operação. As avaliações de processos podem, geralmente, ser efetuadas rapidamente e a um custo razoável. Nos pilotos e nas etapas iniciais de um programa, elas podem ser uma fonte valiosa de informação sobre como melhorar a implementação do programa.

Notas

1. Vide Fiszbein e Schady (2009) para uma visão geral dos programas de CCT e o papel influente desempenhado pelo Progresa/Oportunidades devido à sua avaliação de impacto.
2. Para uma discussão detalhada da análise custo-benefício, vide Belli et al. 2001; Boardman et al. 2001; Brent 1996; ou Zerbe e Dively 1994.

Referências

Bamberger, Michael, Rao & Woolcock. (2010). Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development. *Policy Research Working Paper 5245*. Washington, DC: Banco Mundial.

Behrman, J. & Hoddinott, J. (2001). An Evaluation of the Impact of PROGRESA on Pre-school Child Height. *FCND Briefs 104*. Washington, DC: Instituto Internacional de Pesquisa Política em Alimentos.

Belli, P., Anderson, J., Barnum, H, Dixon, J. & Tan, J. (2001). *Handbook of Economic Analysis of Investment Operations*. Washington, DC: Banco Mundial.

Boardman, A., Vining, A., Greenberg, D. & Weimer, D. (2001). *Cost-Benefit Analysis: Concepts and Practice*. New Jersey: Prentice Hall.

Brent, Robert. 1996. *Applied Cost-Benefit Analysis*. England: Edward Elgar.

Fiszbein, A. & Schady, N. (2009). *Conditional Cash Transfer, Reducing Present and Future Poverty*. Relatório de Pesquisa Política do Banco Mundial. Washington, DC: Banco Mundial.

Gertler, Paul J. 2004. Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment. *American Economic Review* 94 (2): 336–41.

Gertler, Paul J. & Molyneaux, J.M. (1994). How Economic Development and Family Planning Programs Combined to Reduce Indonesian Fertility. *Demography* 31 (1): 33–63.

——. (2000). The Impact of Targeted Family Planning Programs in Indonesia. *Population and Development Review* 26: 61–85.

Imas, L. & Rist, R. (2009). *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington, DC: Banco Mundial.

Kremer, M. & Miguel, E. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72 (1): 159–217.

Kremer, M. Moulin, S. & Namunyu, R. (2003). Decentralization: A Cautionary Tale. Poverty Action Lab Paper 10, Cambridge, MA: Instituto de Tecnologia de Massachusetts.

Levy, S. & Rodríguez, E. (2005). *Sin Herencia de Pobreza: El Programa Progresa-Oportunidades de México*. Washington, DC: Banco de Desenvolvimento Interamericano.

McKay, H., McKay, A., Siniestra, L., Gomez, H. & Lloreda, P. (1978). Improving Cognitive Ability in Chronically Deprived Children. *Science* 200 (21): 270–78.

Poverty Action Lab. 2005. Primary Education for All. *Fighting Poverty: What Works?* 1 (Fall): n.p., disponível em: http://www.povertyactionlab.org.

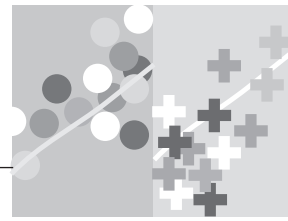
Rao, V. & Woolcock, M. (2003). Integrating Qualitative and Quantitative Approaches in Program Evaluation. In *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, ed. Bourguignon, F.J. & Pereira da Silva, L. 165–90. Nova York: Oxford University Press.

Schultz, P. (2004). School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics* 74 (1): 199–250.

Skoufias, E. & McClafferty, B. (2001). Is *Progresa* Working? Summary of the Results of an Evaluation by IFPRI. Washington, DC: International Food Policy Research Institute.

Vermeersch, C. & Kremer, M. (2005). School Meals, Educational Achievement and School Competition: Evidence from a Randomized Evaluation. *Documentos de Trabalho de Relatório Político 3523*. Washington, DC: Banco Mundial.

Zerbe, R., & Dively, D. (1994). *Benefit Cost Analysis in Theory and Practice*. Nova York: Harper Collins Publishing.



CAPÍTULO 2

Determinando as Perguntas de Avaliação

Este capítulo define os passos iniciais na elaboração de uma avaliação. Os passos incluem a definição do tipo de pergunta a ser respondida pela avaliação, a construção da teoria de mudança, que detalha como o projeto deve alcançar os resultados pretendidos, o desenvolvimento de uma cadeia de resultados, a formulação das hipóteses a serem testadas pela avaliação e a seleção dos indicadores de desempenho.

Todos estes passos contribuem para determinar a pergunta de avaliação e devem ser dados no início do programa, envolvendo uma gama de atores relevantes - dos formuladores de políticas aos administradores do programa - com o objetivo de estabelecer uma visão comum dos objetivos do programa e de como eles serão atingidos. Tal envolvimento gera consensos em relação às principais questões a serem respondidas e fortalece os laços entre a avaliação, a execução do programa e a política. A aplicação dos passos fornece clareza e especificidade, úteis tanto para o desenvolvimento de uma boa avaliação de impacto como para o desenho e implementação de um programa efetivo. Cada passo - desde a clara especificação de objetivos e perguntas, passando pela articulação de ideias embutidas na teoria de mudança, até os resultados que o programa espera gerar - é claramente definido e articulado dentro do modelo lógico expresso na cadeia de resultados.

Tipos de Perguntas de Avaliação

Toda avaliação começa com a formulação de uma pergunta de estudo, que dá foco à pesquisa e que é ajustada de acordo com o interesse da política em questão. Assim, avaliação consiste em gerar evidências confiáveis para que essa pergunta seja respondida. Como explicaremos abaixo, a pergunta básica de uma avaliação de impacto pode ser formulada tal como: *qual é o impacto ou efeito causal do programa sobre os resultados de interesse?* Em um exemplo que aplicaremos em toda a Parte 2, a questão de estudo é: *qual é o efeito do programa de Subsídio ao Seguro Saúde sobre os gastos das famílias com saúde?* A pergunta também pode ser orientada para testar opções de programa: *que combinação de campanhas pelo correio e orientação familiar funciona melhor para estimular o aleitamento materno como única fonte de alimentação?* Uma questão clara de avaliação é o ponto de partida de qualquer avaliação efetiva.

Teorias de Mudança

Uma teoria de mudança é a descrição de como uma intervenção é pensada para gerar os resultados desejados. Ela descreve a lógica causal de como e por que um projeto, programa ou política alcançará os resultados pretendidos. A teoria de mudança é a chave subjacente de qualquer avaliação de impacto, dado o enfoque de causa e efeito da pesquisa. Como um dos primeiros passos de um desenho de avaliação, uma teoria de mudança pode ajudar a especificar as questões da pesquisa.

As teorias de mudança descrevem uma sequência de eventos que levam aos resultados; elas exploram as condições e pressupostos necessários para que a mudança ocorra, tornam explícita a lógica causal por trás do programa e mapeiam as intervenções do programa ao longo de trajetórias lógico-causais. Trabalhar com os atores relevantes do programa para construir uma teoria de mudança pode esclarecer e melhorar o desenho do programa. Isso é especialmente importante em programas que procuram influenciar o comportamento: teorias de mudança podem ajudar a desvincular os insumos e as atividades que compõem as intervenções do programa, os produtos a serem entregues e os resultados que derivam das mudanças comportamentais esperadas entre os beneficiários.

O melhor momento para desenvolver uma teoria de mudança para o programa é no início do processo de desenho, quando as partes interessadas podem reunir-se para desenvolver uma visão comum para o programa, seus objetivos e os caminhos para que esses objetivos sejam atingidos. As partes

interessadas podem, então, começar a execução do programa a partir de um entendimento comum do mesmo, de seu funcionamento e dos objetivos.

Além disso, os formuladores do programa devem revisar a literatura para buscar relatos de experiências sobre programas semelhantes e devem verificar os contextos e pressupostos por trás dos caminhos causais na teoria de mudança que eles estabelecem. Por exemplo, no caso do projeto de chão de concreto, no México, descrito no Quadro 2.1, a literatura fornecerá

Quadro 2.1: *Teoria de Mudança*

Do chão de cimento à felicidade no México

Na sua avaliação do projeto "Piso Firme" (em espanhol) ou "Chão Firme", Cattaneo et al. (2009) examinou o impacto da melhoria da moradia na saúde e bem-estar. Tanto o projeto quanto a avaliação foram motivados por uma teoria clara de mudança.

O objetivo do projeto "Piso Firme" é melhorar os padrões de vida, especialmente a saúde dos grupos vulneráveis vivendo em regiões de baixa renda e densamente povoadas do México. O projeto começou no estado de Coahuila, localizado ao norte do país, e se baseava em uma avaliação situacional realizada pelo governador Enrique Martínez e sua equipe de campanha.

A cadeia de resultados do programa é clara. Foi feita uma sondagem de porta em porta nos bairros elegíveis e foram oferecidos às famílias até 50 metros quadrados de cimento. O governo compra e entrega o cimento e as famílias e os voluntários da comunidade contribuem com a mão-de-obra para instalar o chão. O produto é a construção do chão de cimento, que pode ser concluída em cerca de um dia. Os resultados esperados da melhoria do ambiente doméstico incluem limpeza, saúde e felicidade.

A lógica desta cadeia de resultados é que os chãos de terra batida são um vetor de parasitas, porque este tipo de chão dificilmente se mantém limpo. Os parasitas vivem e se reproduzem em fezes e podem ser ingeridos pelas pessoas quando para casa por animais ou nos calçados. As evidências demonstram que crianças pequenas que moram em casas com chãos de terra têm maior tendência de ser infectadas por parasitas intestinais, que podem causar diarreia e desnutrição, geralmente levando a problemas no desenvolvimento cognitivo ou até mesmo à morte. Chãos de cimento interrompem a transmissão de infestações parasitárias. Eles também permitem um melhor controle de temperatura e são esteticamente mais agradáveis.

Os resultados esperados informaram as perguntas da pesquisa abordadas na avaliação por Cattaneo e seus colegas. Eles esperavam concluir que substituir o chão de terra por chão de cimento reduziria a incidência de diarreia, desnutrição e deficiência de micronutrientes. Isto, por sua vez, resultaria em um melhor desenvolvimento cognitivo das crianças. Os pesquisadores também anteciparam e fizeram testes sobre melhorias no bem-estar dos adultos, conforme medido pela maior satisfação das pessoas com a situação de seu domicílio e baixas taxas de depressão e estresse.

Fonte: Catteneo et al. 2009.

informações valiosas sobre como os parasitas são transmitidos e como a infestação de parasitas leva à diarreia infantil.

A Cadeia de Resultados

Uma teoria de mudança pode ser modelada de várias formas - por exemplo, usando modelos teóricos e lógicos, estruturas lógicas, modelos de resultados e cadeias de resultados[1]. Tudo isso inclui os elementos básicos de uma teoria de mudança, isto é, uma cadeia causal, condições externas e influências e pressupostos-chave. Neste livro, usaremos o modelo da cadeia de resultados, pois achamos que é o modelo mais simples e mais claro para descrever a teoria de mudança no contexto operacional dos programas de desenvolvimento.

Conceito-chave:
Uma cadeia de resultados estabelece a sequência de insumos, atividades e produtos que são esperados para melhorar os resultados e o resultado final.

Uma cadeia de resultados estabelece uma linha plausível e lógica de como uma sequência de insumos, atividades e produtos pelos quais um projeto é diretamente responsável interage com o comportamento para estabelecer trajetos através dos quais os impactos são atingidos (Figura 2.1). Ela estabelece a lógica causal desde o início do projeto, começando com os recursos disponíveis, até o final, analisando os objetivos de longo prazo. Uma cadeia básica de resultados mapeará os seguintes elementos:

Insumos: recursos à disposição do projeto, incluindo equipe e orçamento.

Atividades: medidas tomadas ou trabalho executado para converter os insumos em produtos

Produtos: bens e serviços tangíveis que as atividades do projeto produzem (eles ficam diretamente sob o controle do órgão responsável pela implementação.)

Resultados: os resultados prováveis de serem atingidos, uma vez que a população beneficiada use os produtos do projeto (geralmente, são atingidos no curto-médio prazo.)

Resultados finais: os objetivos finais do projeto (podem ser influenciados por vários fatores e são, tipicamente, atingidos em um período de tempo mais longo.)

A cadeia de resultados possui três partes principais:

Implementação: o trabalho que, planeja-se, será fornecido pelo projeto/programa, incluindo insumos, atividades e produtos. Estas são as áreas que o órgão responsável pela implementação pode monitorar diretamente para medir o desempenho do projeto.

Figura 2.1 O que é uma Cadeia de Resultados?

Fonte: autores, com base em várias fontes.

Resultados: os resultados esperados consistem nos resultados imediatos e resultados finais, que não estão sob o controle direto do projeto e são contingentes às mudanças comportamentais dos beneficiados do programa. Em outras beneficiários do programa. Em outras palavras, eles dependem das interações entre o lado da oferta (implementação) e o lado da demanda (beneficiários). Estas são as áreas sujeitas à avaliação de impacto para medir a efetividade.

Pressupostos e riscos: estes não estão representados na Figura 2.1. Incluem quaisquer evidências da literatura sobre a lógica causal proposta e os pressupostos dos quais ela depende, referências ao desempenho de programas semelhantes e uma menção sobre os riscos que podem afetar a obtenção dos resultados pretendidos, bem como qualquer estratégia de mitigação proposta para gerir tais riscos.

Por exemplo, imagine que o ministro da educação do país *A* esteja pensando em introduzir uma nova abordagem no ensino de matemática no ensino médio. Conforme apresentado na Figura 2.2, os insumos do programa

Figura 2.2 Cadeia de Resultados de um Programa de Matemática no Ensino Médio

Fonte: autores através de várias fontes.

incluiriam a equipe do ministério, os professores do ensino médio, um orçamento para o novo programa de matemática e as instalações municipais onde os professores de matemática serão treinados. As atividades do programa consistem na elaboração do novo currículo de matemática, além do desenvolvimento de um programa de treinamento para os professores, o próprio treinamento para os professores, o preparo, impressão e distribuição de novos livros didáticos. Os produtos são o número de professores treinados, a quantidade de livros entregues em salas de aula e a adaptação de testes padronizados ao novo currículo. Os resultados de curto prazo consistem no uso dos novos métodos e livros nas salas de aula e a aplicação dos novos testes padronizados. Os resultados de médio prazo são melhorias no desempenho dos alunos nos testes padronizados de matemática. Os resultados finais são maiores taxas de conclusão do ensino médio e taxas mais altas de empregabilidade e maiores salários para os egressos.

As cadeias de resultados são úteis para todos os projetos, independentemente de se incluirão ou não uma avaliação de impacto, pois permitem que os legisladores e gestores de programas tornem explícitos os objetivos do programa, ajudando-os, portanto, a entender a lógica causal e a sequência de eventos por trás de um programa. As cadeias de resultados também facilitam as discussões sobre o monitoramento e a avaliação, ao evidenciar quais informações precisam ser monitoradas e quais mudanças nos resultados precisam ser incluídas quando o projeto for avaliado.

Para comparar abordagens alternativas do programa, as cadeias de resultados podem ser agregadas em árvores de resultados representando todas as opções viáveis e consideradas durante o desenho do programa ou sua reestruturação. Essas árvores de resultados representam alternativas operacionais e de política para atingir objetivos específicos; elas podem ser usadas ao considerar quais opções de programas poderiam ser testadas e avaliadas. Por exemplo, se o objetivo é melhorar a educação financeira, podem ser investigadas alternativas como uma campanha publicitária versus a instrução para adultos em sala de aula.

Hipóteses para a Avaliação

Uma vez que você tenha descrito a cadeia de resultados, as hipóteses a serem testadas usando a avaliação de impacto podem ser formuladas. No exemplo do ensino de matemática para alunos do ensino médio, as hipóteses a ser testadas poderiam ser:

- O novo currículo é superior ao antigo, no que tange a comunicação de conhecimentos de matemática.
- Professores treinados usam o novo currículo de uma forma mais efetiva que os outros professores.
- Se treinarmos os professores e distribuirmos os livros, então os professores usarão os novos livros e o currículo em sala de aula e os alunos acompanharão o currículo.
- Se treinarmos os professores e distribuirmos os livros, então os resultados dos testes de matemática aumentarão, em média, cinco pontos.
- O desempenho em matemática no ensino médio influencia as taxas de conclusão e o desempenho no mercado de trabalho.

Selecionando Indicadores de Desempenho

Conceito-chave:
Bons indicadores são específicos, mensuráveis, atribuíveis, realistas e direcionados (tem foco).

Uma cadeia de resultados claramente articulada fornece um mapa útil para selecionar os indicadores que serão medidos ao longo da cadeia. Eles incluirão indicadores usados tanto para monitorar a implementação do programa quanto para avaliar os resultados. Mais uma vez, é útil envolver as partes interessadas no programa na seleção destes indicadores, para garantir que os indicadores selecionados sejam boas medidas do desempenho

do programa. O acrônimo *SMART* é amplamente usado como regra útil e geral para garantir que o indicador usado seja:

- *Específico:* para medir a informação que, de fato, se quer medir.
- *Mensurável:* para garantir que as informações sejam, de fato, obtidas.
- *Atribuível:* para garantir que cada medida seja relacionada aos esforços do projeto.
- *Realista:* para garantir que os dados possam ser obtidos em tempo adequado, com frequência razoável e a um custo também razoável.
- *Tem foco (direcionado):* a população-alvo.

Ao escolher indicadores, lembre-se de que é importante identificar indicadores em toda a cadeia de resultados e não somente nos resultados finais, de modo que você possa rastrear a lógica causal de qualquer resultado observado no programa. Mesmo quando você realiza uma avaliação de impacto, é importante rastrear indicadores de implementação, para que você possa determinar se as intervenções estão sendo executadas conforme planejadas, se atingiram os beneficiados pretendidos e se chegaram a tempo (vide Kusek e Rist 2004 ou Imas e Rist 2009 para discussões de como selecionar indicadores de desempenho). Sem estes indicadores ao longo da cadeia de resultados, a avaliação de impacto produzirá somente uma "caixa preta" que identifica se os resultados previstos se materializaram ou não, mas não será capaz de explicar a razão do sucesso ou do fracasso.

Além de selecionar os indicadores, é útil também considerar os arranjos para a produção de dados. A Tabela 2.1 lista os elementos básicos de um plano de monitoramento e avaliação (M&A), cobrindo os arranjos necessários para produzir cada um dos indicadores de modo confiável e tempestivo.

Roteiro para as Partes 2 e 3

Nesta primeira parte do livro, discutimos por que uma avaliação de impacto deveria ser realizada e quando vale a pena realizá-la. Revisamos os vários objetivos que uma avaliação de impacto pode atingir e destacamos as questões de políticas fundamentais que uma avaliação pode abordar. Insistimos na necessidade de se traçar cuidadosamente a teoria de mudança que explica os canais por meio dos quais um programa pode influenciar os resultados finais. Essencialmente, as avaliações de impacto testam se, na prática, a teoria de mudança funciona ou não.

Tabela 2.1 Elementos de um Plano de Monitoramento e Avaliação

Elemento	Descrição
Resultados esperados (resultados imediatos e produtos)	Obtidos dos documentos de elaboração do programa e da cadeia de resultados.
Indicadores (com linhas de base e metas indicativas)	Derivados da cadeia de resultados; os indicadores devem ser *SMART*.
Dados das fontes	Fonte e localização de onde os dados serão obtidos - por exemplo, uma pesquisa, uma revisão, uma reunião de atores relevantes.
Frequência de dados	Frequência de disponibilidade de dados.
Responsabilidades	Quem é responsável pela organização da coleta de dados e a verificação da qualidade e da fonte dos dados?
Análise e relatório	Frequência de análise, método de análise e responsabilidade pelo relatório.
Recursos	Estimativa de recursos requeridos e comprometidos com a execução das atividades M&A planejadas de M&A.
Uso final	Quem receberá e revisará as informações? A que propósito elas servem?
Riscos	Quais são os riscos e pressupostos na execução das atividades M&A planejadas de M&A? Como eles poderiam afetar os eventos de M&A e a qualidade dos dados?

Fonte: Adaptado do UNDP 2009.

Na Parte 2, consideramos *como avaliar*, revisando várias metodologias alternativas que produzem grupos de comparação válidos e permitem a estimativa de impactos válidos do programa. Começamos a introduzir o *cenário contrafactual* como o cerne de qualquer avaliação de impacto, detalhando as propriedades que a estimativa do cenário contrafactual deve ter e fornecendo exemplos de estimativas inválidas ou falsas do mesmo. Apresentamos, então, opções de avaliação de impacto que podem produzir estimativas válidas do contrafactual. Em particular, discutimos a intuição básica por trás de quatro categorias de metodologias: métodos de *seleção aleatória, desenho de regressão descontínua, diferenças-em-diferenças* e *método de pareamento*. Discutimos por que e como cada método pode produzir uma estimativa válida do contrafactual, em que contexto de política cada um pode ser usado e as principais limitações de cada método. Ao longo desta parte do livro, um estudo de caso - o Programa de Subsídio ao Seguro Saúde[2] - é usado para

ilustrar como os métodos podem ser aplicados. Além disso, apresentamos exemplos específicos de avaliações de impacto que fizeram uso dos diferentes métodos.

A Parte 3 delineia os passos para implementar, administrar ou contratar uma avaliação de impacto. Presumimos que, a essa altura, os objetivos da avaliação já tenham sido definidos, a teoria de mudança tenha sido formulada e as perguntas da avaliação tenham sido especificadas. Revisamos as questões-chave que precisam ser respondidas ao formular-se um plano de avaliação de impacto. Começamos fornecendo regras claras para decidir de onde virão os grupos de comparação. É estabelecida uma estrutura simples para determinar quais das metodologias de avaliação de impacto apresentadas na Parte 2 são as mais adequadas para um determinado programa, dependendo de suas regras operacionais. Revisamos, então, os passos nas quatro fases de realização de uma avaliação: operacionalizar a avaliação planejada, escolher uma amostra, coletar dados e produzir e disseminar os resultados.

Notas

1. A University of Wisconsin-Extension (2010) contém uma discussão detalhada sobre como construir uma cadeia de resultados, assim como uma lista abrangente de referências. Imas e Rist (2009) fornecem uma boa revisão de teorias de mudança.
2. NOTA DO TRADUTOR: o termo "seguro saúde" não é comum no Brasil devido à existência do Sistema Único de Saúde (SUS), que oferece serviços de saúde gratuitos à população carente.

Referências

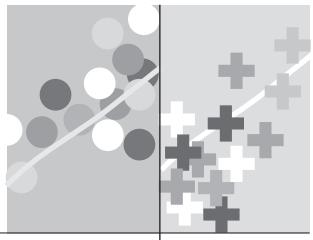
Cattaneo, M., Galiani, Gertler, Martinez e Titiunik. 2009. Housing, Health and Happiness. *American* Economic *Journal: Economic Policy* 1 (1): 75–105.

Imas, L. & Rist, R. (2009). *The Road to Results: Designing and Conducting Effective Development Evaluations*. Washington, DC: Banco Mundial.

Kusek, Z. & Rist. (2004). *Ten Steps to a Results-Based Monitoring and Evaluation System*. Washington, DC: Banco Mundial.

PNUD, Programa das Nações Unidas para o Desenvolvimento. (2009). *Handbook on Planning, Monitoring and Evaluating for Development Results*. Nova York: PNUD.

Universidade de Wisconsin-Extensão. (2010). Enhancing Program Performance with Logic Models. Curso on-line. Disponível em: http://www.uwex.edu/ces /pdande/evaluation/ evallogicmodel.html.



Parte 2

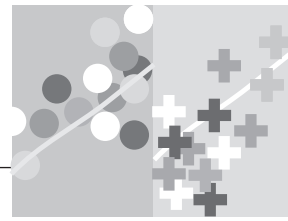
COMO AVALIAR

Agora que estabelecemos as razões por que avaliar o impacto dos programas e das políticas, a segunda parte deste livro explica o que as avaliações de impacto fazem, a quais perguntas respondem, que métodos estão disponíveis para realizá-las e as vantagens e desvantagens de cada um. O menu de opções de avaliações de impacto discutido inclui métodos de seleção aleatórios, regressão descontínua, diferença-em-diferenças e pareamento.

Conforme discutimos na primeira parte, uma avaliação de impacto procura estabelecer e quantificar como uma intervenção afeta os resultados de interesse para os analistas e gestores de políticas. Nesta parte, apresentaremos e examinaremos um estudo de caso, o "Programa de Subsídio ao Seguro Saúde" (HISP). Responderemos à mesma pergunta de avaliação em relação ao HISP diversas vezes, usando a mesma fonte de dados, mas obtendo respostas diferentes (e, às vezes, conflitantes), dependendo da metodologia usada (o leitor deverá assumir que os dados foram adequadamente limpos,

para eliminar qualquer problema relacionado aos dados). A sua tarefa será determinar por que a estimativa de impacto do HISP muda com cada método e quais resultados você considera suficientemente confiáveis para servir de base para a recomendação de políticas.

O caso da HISP é um exemplo de um governo que executa uma reforma de grande escala no setor da saúde, com o objetivo de melhorar a saúde da população. Dentro do objetivo geral, a reforma visa aumentar o acesso à saúde e melhorar a qualidade dos serviços de saúde nas zonas rurais, para elevá-los aos padrões e cobertura que prevalecem nas zonas urbanas. O inovador (e potencialmente oneroso) programa HISP está sendo pilotado. O programa subsidia o seguro saúde para famílias pobres da zona rural, cobrindo custos relacionados aos cuidados primários de saúde e medicamentos. O objetivo central do HISP é reduzir o custo dos cuidados com a saúde incorridos pelas famílias pobres, e, então, melhorar os resultados na saúde. Os gestores estão considerando expandir o HISP para o país inteiro. Expandir o programa custaria centenas de milhões de dólares, mas os gestores estão preocupados que as famílias de baixa renda de zonas rurais não tenham recursos para acessar os serviços básicos de saúde sem subsídio, o que teria consequências prejudiciais à saúde. A pergunta chave da avaliação é: *qual é o efeito do HISP nos gastos das famílias com saúde e no estado de saúde das famílias pobres?* Respostas a perguntas como estas orientam os gestores na decisão de que políticas adotar e quais programas implementar. Estas políticas e programas, por sua vez, podem afetar o bem-estar de milhões de pessoas em todo o mundo. Esta parte do livro discutirá como responder rigorosamente a estas perguntas fundamentais da avaliação.



CAPÍTULO 3

Inferência Causal e Cenário Contrafactual

Começamos por examinar dois conceitos que se integram ao processo de condução de avaliações que sejam rigorosas e confiáveis: inferência causal e contrafatuais.

Inferência Causal

A questão básica da avaliação de impacto constitui, essencialmente, um problema de inferência causal. Avaliar o impacto de um programa em uma série de resultados equivale a avaliar o efeito causal do programa sobre estes resultados. A maioria das questões sobre políticas envolvem relação de causa e efeito: a formação do professor melhora os resultados das provas dos alunos? Os programas de transferência de renda geram resultados melhores na saúde das crianças? Os programas de capacitação profissional aumentam a renda dos indivíduos capacitados?

Embora as questões de causa e efeito sejam comuns, não é uma questão simples determinar se a relação é causal. No contexto de um programa de capacitação profissional, por exemplo, simplesmente observar que a renda de um indivíduo capacitado aumenta após ele ou ela ter concluído tal programa não é suficiente para estabelecer a causalidade. A renda do indivíduo

capacitado poderia ter aumentado mesmo se ele não tivesse participado de nenhum treinamento, devido a seus próprios esforços, a mudanças nas condições do mercado de trabalho, ou mesmo por causa de uma série de outros fatores que possam afetar a renda. As avaliações de impacto nos ajudam a superar o desafio de determinar a causalidade ao demonstrar empiricamente em que grau um determinado programa - e somente este programa - contribuiu para a mudança de um resultado. Para estabelecer causalidade entre um programa e um resultado, usamos métodos de avaliação de impacto para descartar a possiblidade de que quaisquer outros fatores, além do programa de interesse, expliquem o impacto observado.

A resposta à pergunta básica da avaliação de impacto - *qual é o impacto ou o efeito causal de um programa P sobre um resultado de interesse Y?* - é dada pela fórmula básica da avaliação de impacto:

$$\alpha = (Y \mid P = 1) - (Y \mid P = 0)$$

De acordo com esta fórmula, o impacto causal (α) de um programa (P) sobre um resultado (Y) corresponde à diferença entre o resultado (Y) na presença do programa (em outras palavras, quando $P = 1$) e o mesmo resultado (Y) na ausência do programa (isto é, quando $P = 0$).

Por exemplo, se P denota um programa de capacitação profissional e Y designa renda, então, o impacto causal do programa de treinamento vocacional (α) é expresso pela diferença entre a renda do indivíduo (Y) após participar do programa de treinamento vocacional (em outras palavras, quando $P = 1$) e a renda do mesmo indivíduo (Y), no mesmo momento no tempo, considerando que este não tenha participado do programa (ou seja, quando $P = 0$). Em outras palavras, gostaríamos de medir a renda no mesmo instante no tempo, para a mesma unidade de observação (um indivíduo, neste caso), mas em dois cenários diferentes. Se fosse possível fazer isso, estaríamos observando qual seria o valor da renda do mesmo indivíduo no mesmo ponto do tempo com e sem a participação no programa, de forma que a *única* explicação possível para qualquer diferença na renda do indivíduo seria o programa. Ao comparar o indivíduo com ele mesmo, no mesmo momento, seriam eliminados quaisquer fatores exógenos que poderiam igualmente ter explicado a diferença nos resultados. Poderíamos, então, dizer com segurança que a relação entre o programa de capacitação profissional e a renda é causal.

A fórmula básica da avaliação de impacto é válida para qualquer unidade que estiver sendo analisada: um indivíduo, uma família, uma comunidade, uma empresa, uma escola, um hospital, ou qualquer outra unidade de observação que possa receber ou ser impactada por um programa. A fórmula também é válida para qualquer resultado (Y) que esteja plausivelmente relacionado ao

programa em questão. Uma vez medidos os dois componentes-chave desta fórmula - o resultado (Y), tanto na presença quanto na ausência do programa - poderemos responder a qualquer pergunta sobre o impacto do programa.

O Contrafactual

Conforme discutido anteriormente, podemos pensar no impacto (α) de um programa como a diferença nos resultados (Y) para o mesmo indivíduo, com e sem a sua participação em um determinado programa. Contudo, sabemos que é impossível medir o mesmo indivíduo em duas situações diferentes ao mesmo tempo. Em um dado instante no tempo, um indivíduo participou ou não de um dado programa. O indivíduo não pode ser observado simultaneamente em dois estados diferentes (em outras palavras, participando e não participando de um programa). Isto se chama o "problema do contrafactual": Como é possível medir o que teria acontecido se a outra circunstância tivesse prevalecido? Embora possamos observar e medir o resultado (Y) para os participantes do programa ($Y \mid P = 1$), não há informações para estabelecer quais teriam sido os resultados na ausência do programa ($Y \mid P = 0$). Na fórmula básica da avaliação de impacto, o termo ($Y \mid P = 0$) *representa o cenário contrafactual*. Podemos ver isto como *o que teria acontecido* se um indivíduo não tivesse participado do programa. Em outras palavras, o contrafactual é qual teria sido o resultado (Y) na ausência do programa (P).

Conceito-chave:
O contrafactual é uma estimativa do que seria o resultado (Y) para um participante do programa na ausência deste programa (P).

Por exemplo, imagine que o "Sr. Desafortunado" tome um comprimido vermelho e morra cinco dias depois. Somente porque o Sr. Desafortunado morreu após ter tomado o comprimido vermelho não se pode concluir que o comprimido vermelho tenha *provocado a* sua morte. Talvez ele estivesse muito doente quando tomou o comprimido vermelho e tenha sido a doença, e não o comprimido vermelho, que provocou a sua morte. Inferir causalidade exige a exclusão de outros fatores potenciais que possam afetar os resultados em consideração. No simples exemplo de determinar se o fato do Sr. Desafortunado ter tomado o comprimido vermelho provocou a sua morte, um avaliador precisa estabelecer o que teria acontecido com o Sr. Desafortunado se ele *não* tivesse tomado o comprimido. Na medida em que o Sr. Desafortunado de fato tomou o comprimido vermelho, não é possível observar diretamente o que teria acontecido caso ele o não tivesse feito. O que teria acontecido com ele caso não tivesse tomado o comprimido vermelho é o cenário contrafactual e o principal desafio do avaliador é determinar como tal estado contrafactual da realidade se manifestaria (vide quadro 3.1).

Ao realizar uma avaliação de impacto, é relativamente simples obter-se o primeiro termo da fórmula básica ($Y \mid P = 1$), o resultado do tratamento.

Simplesmente medimos o resultado de interesse para a população que participou do programa. No entanto, o segundo termo da fórmula $(Y \mid P = 0)$ não pode ser observado diretamente pelos participantes do programa - por isso a necessidade de preencher esta lacuna de informação *estimando-se o contrafactual*. Para tanto, usamos, geralmente, *grupos de comparação*

Quadro 3.1: Estimativa do Contrafactual

A Srta. Única e o Programa de Transferência de Renda

A Srta. Única é uma garotinha recém-nascida, cuja mãe receberá uma transferência de renda mensal contanto que a Srta. Única faça exames médicos regulares no posto de saúde local, seja vacinada e tenha seu crescimento monitorado. O governo supõe que a transferência de renda motivará a mãe da Srta. Única a buscar os serviços de saúde requeridos pelo programa e ajudará a Srta. Única a crescer forte e com a altura adequada. Para a avaliação de impacto, o governo escolheu o crescimento como um indicador de resultado para saúde no longo prazo e medirá a altura da Srta. Única após três anos de duração do programa de transferência de renda.

Suponha que você possa medir a altura da Srta. Única aos três anos de idade. Idealmente, para avaliar o impacto do programa, você teria que medir a altura da Srta. Única aos três anos de idade, considerando que sua mãe tenha recebido a transferência de renda e, também, a altura da Srta. Única aos três anos de idade caso sua mãe não tivesse recebido a transferência de renda. Você compararia, então, as duas alturas. Se você pudesse comparar a altura da Srta. Única aos 3 anos de idade com e sem o programa, você saberia que qualquer diferença na altura teria sido causada somente por conta programa. Porque tudo mais a respeito da Srta. Única seria igual, não haveria outras características que pudessem explicar a diferença na altura.

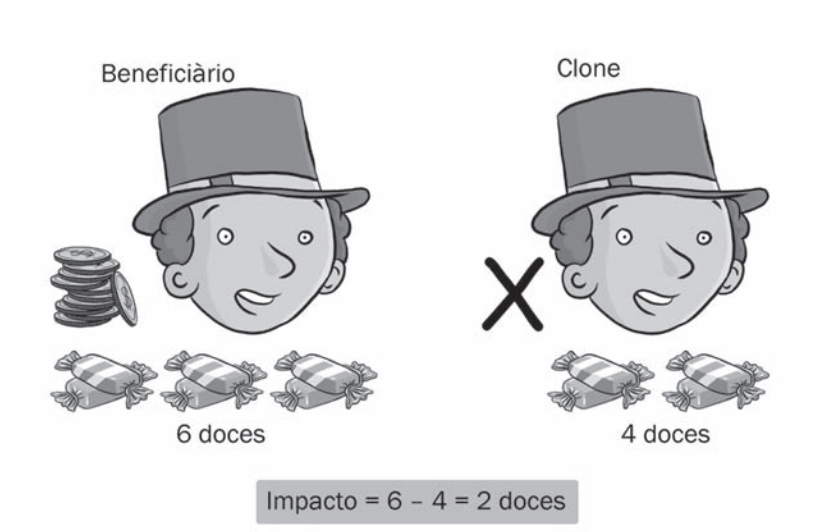
Infelizmente, no entanto, é impossível observar a Srta. Única com e sem o programa de transferência de renda: ou a família dela participa ou não participa do programa. Em outras palavras, não sabemos qual seria o cenário contrafactual. Uma vez que a mãe da Srta. Única de fato participou do programa de transferência de renda, nós não podemos saber qual teria sido a altura da Srta. Única caso sua mãe não tivesse recebido a transferência de renda. Encontrar uma comparação adequada para a Srta. Única será um desafio, precisamente porque a Srta. Única é única. Sua exata situação socioeconômica, características genéticas e pessoais não podem ser encontradas em mais ninguém. Caso estivéssemos simplesmente comparando a Srta. Única a criança que não se inscreveu no programa de transferência de renda, digamos, o Sr. Inigualável, a comparação poderia não ser adequada. A Srta. Única não é idêntica ao Sr. Inigualável. A Srta. Única e o Sr. Inigualável podem não se assemelhar, podem não viver no mesmo local, podem não ter os mesmos pais e podem não ter apresentado a mesma altura ao nascer. Desse modo, se observarmos que o Sr. Inigualável é mais baixo do que a Srta. Única aos 3 anos de idade, não poderemos saber se a diferença é devida ao programa de transferência de renda ou a uma das muitas outras diferenças entre essas duas crianças.

(às vezes denominados “grupos de controle”). O restante da segunda parte deste livro enfocará os diferentes métodos ou abordagens que podem ser usados para identificar grupos válidos de comparação que reproduzam exatamente ou simulem o cenário contrafactual. Identificar tais grupos de comparação é o ponto crucial de toda avaliação de impacto, independentemente do tipo de programa sendo avaliado. Em suma, sem uma estimativa válida do contrafactual, o impacto de um programa não pode ser estabelecido.

Estimando o Contrafactual

A fim de ilustrar melhor a estimação do contrafactual, passaremos a um exemplo hipotético que, embora não seja relevante enquanto política, nos ajudará a refletir um pouco mais sobre este conceito-chave. Em nível conceitual, resolver o problema do contrafactual requer que o avaliador identifique um”clone perfeito” para cada participante (figura 3.1). Por exemplo, digamos que o Sr. Fulano de Tal recebe um adicional de $12 na sua mesada e que queremos medir o impacto deste tratamento em seu consumo de doces. Se fosse possível identificar um clone perfeito para o Sr. Fulano de Tal, a avaliação seria fácil: poderíamos simplesmente comparar o número

Figura 3.1 O Clone Perfeito



Fonte: Autores

de doces consumidos pelo Sr. Fulano de Tal (digamos que seis) ao número de doces consumidos por seu clone (digamos que quatro). Neste caso, o impacto da mesada adicional seria a diferença entre esses dois valores, ou seja, dois doces. Na prática, sabemos que é impossível identificar clones perfeitos: mesmo entre gêmeos geneticamente idênticos há diferenças importantes.

Embora não exista um clone perfeito para um único indivíduo, existem ferramentas estatísticas que podem ser usadas para gerar dois grupos de indivíduos que, se forem grandes o suficiente, serão estatisticamente indistinguíveis um do outro. Na prática, um objetivo-chave de uma avaliação de impacto é identificar um grupo de participantes do programa (ou grupo de tratamento) e um grupo de não-participantes (o grupo de comparação), que sejam estatisticamente idênticos na ausência do programa. Se os dois grupos forem idênticos, com a única exceção de que um grupo participa do programa e o outro não, então poderemos ter a certeza de que qualquer diferença nos resultados deve-se ao programa.

Conceito-chave:
Um grupo válido de comparação terá as mesmas características que o grupo de participantes do programa ("grupo de tratamento"), exceto pelo fato de que as unidades no grupo de comparação não se beneficiam do programa.

O principal desafio, assim, é identificar um grupo de comparação válido, que possua as mesmas características que o grupo de tratamento. Especificamente, os grupos de comparação e tratamento devem ser equivalentes em pelo menos três aspectos: Primeiro, o grupo de tratamento e o grupo de comparação devem ser idênticos na ausência do programa. Embora não seja necessário que cada observação dentro do grupo de tratamento seja idêntica a cada observação dentro do grupo de comparação, na média, as características de interesse dos grupos de tratamento e de comparação devem ser as mesmas. Por exemplo, a idade média do grupo de tratamento deve ser a igual à idade média do grupo de comparação. Em segundo lugar, o grupo de tratamento e o grupo de comparação devem apresentar a mesma reação ao programa. Por exemplo, as rendas do grupo de tratamento devem ter a mesma probabilidade de serem beneficiadas pelo treinamento que as rendas do grupo de comparação. Em terceiro lugar, os grupos de tratamento e de comparação não podem ser expostos de maneira diferente a outras intervenções durante o período de avaliação. Por exemplo, se isolarmos o impacto da mesada adicional sobre o consumo de doces, o grupo de tratamento não poderia ser privilegiado também com mais idas à confeitaria do que o grupo de comparação, já que isso poderia confundir os efeitos da mesada com o efeito do maior acesso aos doces.

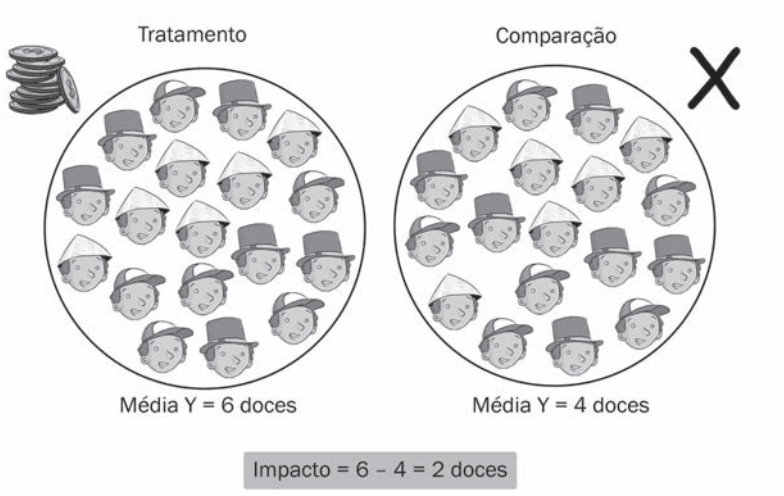
Conceito-chave:
Quando o grupo de comparação para uma avaliação for inválido, então a estimativa do impacto do programa também será inválida, não estimará o verdadeiro impacto do programa. Em termos estatísticos, será "enviesado".

Quando essas três condições são atendidas, então somente a existência do programa de interesse explicará quaisquer diferenças no resultado (Y) entre os dois grupos, contanto que o programa seja implementado. O motivo é que a única diferença entre os grupos de tratamento e de comparação é que os membros do grupo de tratamento receberão o programa, enquanto os

membros do grupo de comparação não o receberão. Quando as diferenças nos resultados puderem ser integralmente atribuídas ao programa, o impacto causal do programa terá sido identificado. Então, ao invés de olhar para o impacto da mesada adicional somente para o Sr. Fulano de Tal, poderíamos olhar para o impacto sobre um grupo de crianças (figura 3.2). Se for possível identificar outro grupo de crianças totalmente semelhante, exceto pelo fato de que elas não recebem o acréscimo de mesada, a estimativa do impacto do programa seria a diferença entre os dois grupos em relação à média de consumo de doces. Consequentemente, se o *grupo tratado* consome uma média de 6 doces por pessoa, enquanto o *grupo de comparação* consome uma média de 4, o impacto médio da mesada adicional sobre o consumo de doces seria 2.

Agora que definimos um grupo de comparação válido, é importante considerar o que aconteceria se decidíssemos continuar com a avaliação sem identificar tal grupo. Intuitivamente, isto agora deveria estar claro: um grupo de comparação inválido é aquele que se diferencia do grupo de tratamento por conta de algum outro aspecto que não somente a ausência do tratamento. Estas diferenças adicionais podem invalidar ou, em termos estatísticos, enviesar a estimativa: não será estimado o impacto real do programa. Mais do que isso, será estimado o efeito do programa combinado ao impacto destas outras diferenças.

Figura 3.2 Um Grupo de Comparação Válido



Fonte: Autores.

Dois Tipos de Estimativas de Impacto

Após estimar o impacto do programa, o avaliador precisa saber como interpretar os resultados. Uma avaliação sempre estima o impacto de um programa comparando os resultados do grupo de tratamento à estimativa do cenário contrafactual obtida de um grupo de comparação válido, usando a equação básica da avaliação de impacto. Dependendo do que o tratamento e o contrafactual de fato representem, a interpretação do impacto de um programa pode variar.

O impacto estimado α é chamado de estimativa da "intenção ao tratamento" (ITT, do inglês, *"intention-to-treat"*), quando a fórmula básica se aplica àquelas unidades às quais o programa foi oferecido, independentemente de terem se inscrito ou não no programa. O estimador ITT é importante para os casos nos quais tentamos determinar o impacto médio de um programa na população-alvo do programa. Por outro lado, o impacto estimado α é chamado de "efeito do tratamento sobre os tratados" (estimador TOT, do inglês, treatment-on-the-treated), quando a fórmula básica se aplica àquelas unidades às quais o programa foi oferecido e que realmente se inscreveram. Os estimadores ITT e TOT serão os mesmos quando houver um cumprimento perfeito, isto é, quando todas as unidades para as quais o programa foi oferecido de fato decidiram inscrever-se nele. Voltaremos a falar sobre a diferença entre os estimadores ITT e TOT em detalhes mais adiante. Começaremos com um exemplo.

Considere o Programa de Subsídio ao Seguro Saúde, ou HISP (*Health Insurance Subsidy Program,* na sigla em inglês), exemplo descrito na introdução da parte 2, no qual qualquer família em uma comunidade de tratamento pode inscrever-se em um subsídio de assistência médica. Embora todas as famílias nas comunidades de tratamento estejam qualificadas a se inscrever no programa, uma fração delas, digamos 10%, pode decidir não se inscrever (talvez porque já possuam um plano de assistência médica fornecido pela empresa onde trabalham ou porque sejam saudáveis e não antecipem a necessidade de cuidados médicos, ou devido a inúmeras outras razões). Neste cenário, 90% das famílias na comunidade de tratamento decidem se inscrever no programa e, de fato, receber os serviços que o programa oferece. O estimador ITT seria obtido computando-se a fórmula básica de avaliação de impacto para todas as famílias às quais o programa foi oferecido, isto é, para 100% das famílias das comunidades de tratamento. Por outro lado, o estimador TOT seria obtido calculando-se a fórmula básica da avaliação de impacto somente para o subconjunto de famílias que realmente decidiu se inscrever no programa, isto é, 90% das famílias inscritas nas comunidades de tratamento.

Duas Estimativas Falsas do Contrafactual

No restante da parte 2 deste livro, discutiremos os diversos métodos que podem ser usados para construir grupos de comparação válidos que permitirão estimar o cenário contrafactual. Antes disso, no entanto, é útil discutir dois métodos comuns, porém altamente arriscados, de construção de grupos de comparação, que podem levar a estimativas inadequadas do cenário contrafactual. Estas duas estimativas "falsas" dos cenários contrafatuais são: (1) comparações *antes-e-depois*, ou *pré-pós*, que comparam os resultados dos participantes antes e após a introdução do programa; e (2) comparações *com-e-sem,* entre unidades que escolheram se inscrever no programa e as que preferiram não fazê-lo.

Cenário Contrafactual Falso 1: Comparando o Antes e o Depois

A comparação *antes-e-depois* tenta determinar o impacto de um programa acompanhando as mudanças nos resultados para os participantes do programa ao longo do tempo. Voltando à fórmula básica da avaliação de impacto, o resultado para o grupo de tratamento ($Y \mid P = 1$) é simplesmente o resultado pós-intervenção. No entanto, o contrafactual ($Y \mid P = 0$) é estimado usando o resultado da pré-intervenção. Em essência, esta comparação assume que, se o programa nunca tivesse existido, o resultado (Y) para os participantes do programa teria sido exatamente o mesmo que a sua situação pré-programa. Infelizmente, na grande maioria dos casos, esta suposição simplesmente não é válida.

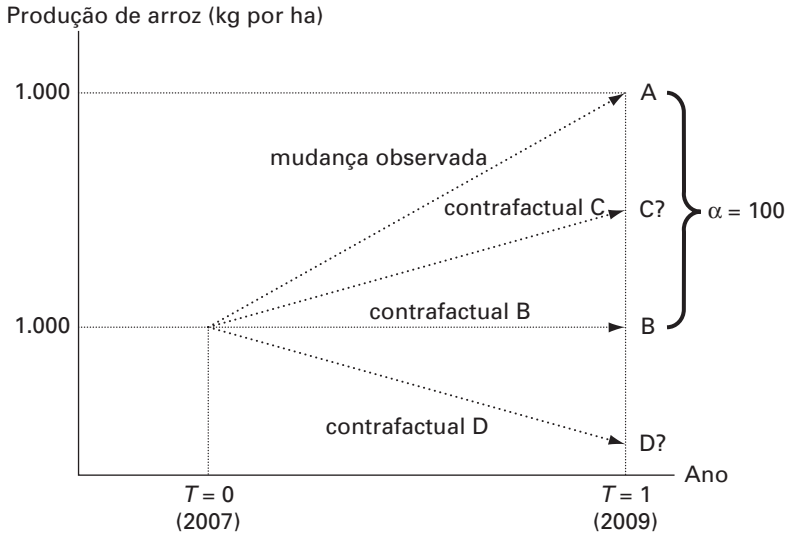
Podemos tomar como exemplo a avaliação de um programa de microfinanças para agricultores pobres na zona rural. Digamos que o programa ofereça aos fazendeiros microempréstimos para a compra de fertilizantes para aumentar sua produção de arroz. Nota-se que, um ano antes do lançamento do programa, os fazendeiros colhiam uma média de 1.000 quilogramas (kg) de arroz por hectare. O programa de microfinanças é lançado e, um ano depois, a produção de arroz aumenta para 1.100 kg por hectare. Se você estiver tentando avaliar o impacto usando uma comparação *antes-e-depois*, o resultado antes da intervenção seria usado como cenário contrafactual. Aplicando a fórmula básica de avaliação de impacto, você concluiria que o programa aumentou a produção de arroz em 100 kg por hectare.

No entanto, imagine que o índice pluviométrico tenha sido normal durante o ano anterior ao lançamento do programa, mas que uma seca ocorreu no ano em que o programa começou. Neste contexto, o resultado pré-intervenção não pode constituir um cenário contrafactual adequado.

A Figura 3.3 ilustra o porquê. Como os agricultores receberam o benefício durante um ano de seca, a sua produção média sem o programa de microempréstimos teria sido menor, no nível D, e não no nível B, como assumido pelo método de comparação *antes-e-depois*. Neste caso, o impacto real do programa é maior do que 100 kg. Por outro lado, se as condições ambientais tivessem melhorado efetivamente com o tempo, a produção contrafactual de arroz poderia ter ficado no nível C, no qual o verdadeiro impacto do programa teria sido inferior a 100 kg. Em outras palavras, a menos que possamos contabilizar estatisticamente o índice pluviométrico e *todos os demais fatores* que possam afetar a produção de arroz ao longo do tempo, nós simplesmente não poderemos calcular o verdadeiro impacto do programa por meio de uma comparação *antes-e-depois*.

Embora as comparações *antes-e-depois* possam ser inválidas na avaliação de impacto, isto não significa que não sejam válidas para outros fins. De fato, os sistemas de informação administrativos de muitos programas geralmente registram dados dos participantes ao longo do tempo. Por exemplo, um sistema de informações sobre gestão educacional pode rotineiramente coletar dados sobre as matrículas de alunos nas escolas onde um programa de merenda escolar esteja sendo executado. Estes dados permitem que os gestores do programa observem se o número de crianças matriculadas na escola está aumentando ao longo do tempo. Essas informações são importantes e

Figura 3.3 Estimativas *Antes-e-Depois* de um Programa de Microfinanciamento



Fonte: Autores, com base no exemplo hipotético do texto.

valiosas para os gestores que planejam e desenvolvem relatórios sobre o sistema educacional. No entanto, estabelecer que o programa de merenda escolar tenha *provocado* a mudança observada na matrícula é algo muito mais desafiador, pois muitos fatores diferentes afetam a matrícula de alunos ao longo do tempo. Desse modo, embora monitorar mudanças nos resultados ao longo do tempo para um grupo de participantes seja extremamente valioso, geralmente não nos permite concluir se (ou quanto) um determinado programa contribuiu para uma melhoria, uma vez que podem existir outros fatores que variam no tempo que podem afetar tal resultado.

Nos exemplos do programa de microfinanciamento e da produção de arroz, vimos que muitos fatores podem afetar a produção ao longo do tempo. Do mesmo modo, muitos fatores podem afetar a maioria dos resultados de interesse para os programas de desenvolvimento. Por este motivo, o resultado pré-programa quase nunca é uma boa estimativa do contrafactual, e é por isso que o chamamos de "falso contrafactual".

Fazendo uma Avaliação *Antes-e-Depois* do Programa Subsídio ao Seguro Saúde

Suponhamos que o HISP seja um novo programa no seu país, que oferece subsídio para a compra de plano de assistência médica às famílias rurais de baixa renda, e que este plano cubra despesas relacionadas ao atendimento básico de saúde e à aquisição de medicamentos. O objetivo do HISP é reduzir os gastos das famílias de baixa renda com despesas de saúde e, no final das contas, melhorar os resultados de saúde. Embora muitos indicadores de resultados possam ser considerados na avaliação do programa, o governo está particularmente interessado em analisar os efeitos do HISP no montante que as famílias de baixa renda gastam em cuidados básicos de saúde e medicamentos, medidos como despesas anuais per capita da família (doravante denominadas "despesas com a saúde").

O HISP representará uma grande fatia do orçamento nacional se for expandido para todo o país - até 1,5% do produto interno bruto (PIB), conforme algumas estimativas. Além disso, o desenvolvimento de um programa dessa natureza envolve grandes complexidades logísticas e administrativas. Por estas razões, foi tomada a decisão, nos níveis mais altos do governo, de introduzir o HISP primeiramente como um programa piloto e, depois, dependendo dos resultados da primeira etapa, expandi-lo gradualmente ao longo do tempo. Com base nos resultados de análises financeiras e de custo-benefício, o Presidente e o seu gabinete anunciaram que, para viabilizar o HISP e estendê-lo nacionalmente, o programa deveria reduzir a média anual da despesa per capita com saúde das famílias rurais de baixa

renda em, pelo menos, $9 em relação ao que gastariam na ausência do programa, e isso deveria acontecer dentro de um período de 2 anos.

O HISP será introduzido em 100 municípios rurais durante a fase inicial do programa piloto. Logo antes do início do programa, o governo do seu país contrata uma empresa de pesquisa para realizar um estudo de linha de base incluindo todas as 4.959 famílias destes municípios. O estudo coleta informações detalhadas sobre cada família, incluindo a composição demográfica, ativos, acesso a serviços médicos e despesas com saúde incorridas no ano anterior. Pouco tempo depois da realização do estudo de linha de base, o HISP é introduzido nos 100 municípios piloto acompanhado de grande divulgação, incluindo eventos comunitários e outras campanhas promocionais para motivar as famílias elegíveis a se inscrever.

Das 4.959 famílias na amostra da linha de base, um total de 2.907 se inscrevem no HISP durante os dois primeiros anos do programa. Durante estes dois anos, o HISP opera com sucesso, de acordo com a maioria das medidas. As taxas de cobertura são altas e as pesquisas mostram que a maior parte das famílias inscritas está satisfeita com o programa. Ao término do período de 2 anos do piloto, uma segunda rodada de dados de avaliação é coletada na mesma amostra de 4.959 famílias.[1]

O Presidente e o Ministro da Saúde deixaram você responsável pela supervisão da avaliação de impacto do HISP e lhe incumbiram de informar se o programa deve ou não ser estendido nacionalmente. A sua pergunta de interesse na avaliação de impacto é: *em quanto o HISP reduziu as despesas com saúde das famílias rurais de baixa renda?* Lembre-se que o que está em jogo é muito importante. Se for concluído que o HISP reduziu as despesas com saúde em cerca de $9 ou mais, ele será estendido nacionalmente. Se o programa não atingir a meta de $9, você não recomendará a expansão.

O primeiro consultor "especialista" em avaliação que você contrata informa que, para estimar o impacto do HISP, é preciso calcular a mudança nas despesas com saúde ao longo do tempo para as famílias que se inscreveram. O consultor argumenta que, pelo fato de o HISP cobrir todos os custos relacionados aos cuidados básicos de saúde e medicamentos, qualquer redução nas despesas ao longo do tempo deve ser, em grande parte, atribuída ao efeito do HISP. Desse modo, usando somente o subconjunto de famílias inscritas, você estima a média das despesas com saúde antes da implementação do programa e dois anos mais tarde. Em outras palavras, você executa uma avaliação *antes-e-depois*. Os resultados são apresentados na tabela 3.1.

Você observa que as famílias que se inscreveram no HISP reduziram as despesas com saúde de $14,40, antes da introdução do HISP, para $7,80 dois anos depois - uma redução de $6,60 (ou 45%) durante o período. Conforme

Tabela 3.1 Caso 1 — Impacto do HISP Usando a Comparação *Antes-e-Depois* (Comparação de Médias)

	Depois	Antes	Diferença	Teste t
Despesas da família com saúde	7,8	14,4	–6,6	–28,9

Fonte: Cálculos dos autores a partir de um conjunto de dados hipotéticos.

indicado pelo valor do teste t), a diferença entre as despesas com saúde antes e depois do programa é *estatisticamente significativa,* isto é, a probabilidade de que o efeito estimado seja estatisticamente igual a zero é muito baixa.

Embora a comparação *antes-e-depois* seja para o mesmo grupo de famílias, você se preocupa com o fato de que outros fatores possam ter se modificado ao longo do tempo e tenham afetado as despesas com saúde. Por exemplo, uma variedade de intervenções na saúde têm sido feitas simultaneamente nos municípios em questão. Alternativamente, algumas mudanças nas despesas familiares podem ser resultado da crise financeira pela qual seu país passou recentemente. Para responder a algumas destas questões, o seu consultor conduz uma *análise de regressão* que controlará os fatores externos. Os resultados são apresentados na tabela 3.2

Neste caso, a regressão linear é das despesas com saúde sobre uma variável binária (0-1) que pode assumir o valor de 0, quando a observação se refere à linha de base, ou 1, quando se refere ao momento pós-intervenção. A regressão linear multivariada também *controla* ou *mantém constantes* outras características observadas para as famílias de sua amostra, incluindo indicadores de riqueza (ativos), composição familiar e assim por diante. Note que a regressão linear simples é equivalente à diferença simples entre o antes e o depois nas despesas com a saúde (uma redução de $6,59). Uma vez controlados outros fatores disponíveis em seus dados, será encontrado um resultado semelhante - uma redução de $6,65.

Tabela 3.2 Caso 1 — Impacto do HISP Usando a Comparação *Antes-e-Depois* (Análise de Regressão)

	Regressão linear	Regressão linear multivariada
Impacto estimado nos gastos da família com saúde	–6,59**	–6,65**
	(0,22)	(0,22)

Fonte: Autores.

Observação: Os erros padrão estão entre parênteses.

** 1% de significância estatística.

PERGUNTA 1

A. Com base nos resultados do caso 1, o HISP deveria ser estendido nacionalmente?
B. Pode-se dizer que esta análise controla todos os fatores que afetam as despesas com saúde ao longo do tempo?

Cenário Contrafactual Falso 2: Comparando os Inscritos aos Não Inscritos

A comparação das unidades que recebem um programa com as unidades que não o recebem ("com-e-sem") constitui outro cenário contrafactual falso. Considere, por exemplo, um programa de capacitação profissional para jovens desempregados. Suponha que, dois anos após o lançamento do programa, uma avaliação tente estimar seu impacto sobre a renda comparando as rendas médias de um grupo de jovens que escolheram se inscrever no programa com aquelas de um grupo que decidiu não se inscrever. Suponha que os resultados demonstrem que os jovens que se inscreveram no programa ganham duas vezes mais do que aqueles que não se inscreveram.

Como esses resultados podem ser interpretados? Neste caso, o cenário contrafactual é estimado com base nas rendas de indivíduos que decidiram não se inscrever no programa. No entanto, os dois grupos de jovens tendem a ser fundamentalmente diferentes. Os indivíduos que escolheram participar podem estar altamente motivados a melhorar sua condição de vida e possivelmente têm grandes expectativas quanto ao retorno proporcionado pela capacitação. Por outro lado, os que optaram por não se inscrever podem ser jovens desmotivados que não esperam se beneficiar desse tipo de programa. É provável que estes dois tipos de jovens tenham desempenhos diferentes no mercado de trabalho e também tenham rendas diferentes, mesmo na ausência do programa de capacitação profissional.

Conceito-chave:
O viés de seleção ocorre quando as razões pelas quais um indivíduo participa de um programa estão correlacionadas com os resultados. Este viés normalmente ocorre quando o grupo de comparação não é elegível para o programa ou decide não participar dele.

Desse modo, o grupo que escolheu não se inscrever não oferece uma boa estimativa do cenário contrafactual. Se uma diferença de renda for observada entre os dois grupos, não poderemos determinar se ela é devida ao programa de capacitação ou às diferenças subjacentes, em termos de motivação e de outros fatores, existentes entre os dois grupos. O fato de indivíduos menos motivados decidirem não se inscrever no programa de treinamento leva, consequentemente, a um viés em nossa avaliação do impacto do programa.[2] Este viés é chamado de "viés de seleção". Neste caso, se os jovens que se inscreveram tivessem obtido rendas mais altas mesmo na ausência do programa, o viés de seleção seria positivo; em outras palavras, nós estaríamos superestimando o impacto do programa de capacitação profissional sobre a renda.

Comparando Indivíduos que Escolheram se Inscrever no Programa de Subsídio ao Seguro Saúde a Aqueles que Decidiram não se Inscrever

Refletindo um pouco mais sobre a comparação *antes-e-depois* com a sua equipe de avaliação, você percebe que ainda há muitos fatores que variam no tempo que podem explicar parte da mudança nas despesas com saúde ao longo do tempo (em particular, o Ministro da Fazenda está preocupado que a recente crise financeira tenha afetado as despesas familiares com saúde e possa explicar a mudança observada). Outro consultor sugere que seria mais apropriado estimar o cenário contrafactual no período pós-intervenção, isto é, dois anos após o início do programa. O consultor observa corretamente que, das 4.959 famílias na amostra da linha de base, somente 2.907 realmente se inscreveram no programa e, então, aproximadamente 41% das famílias na amostra permanecem sem a cobertura do HISP. O consultor argumenta que as famílias da mesma localidade seriam expostas às mesmas intervenções de saúde do lado da oferta e às mesmas condições econômicas locais, de modo que os resultados, no período pós-intervenção, do grupo de famílias não inscritas ajudariam a controlar muitos dos fatores que afetam tanto as famílias inscritas quanto as não inscritas.

Assim, você decide calcular as despesas médias com saúde no período pós-intervenção, tanto para as famílias que se inscreveram no programa quanto para as famílias que decidiram não se inscrever, produzindo as observações apresentadas na tabela 3.3.

Usando as despesas médias com saúde das famílias não inscritas como estimativa do cenário contrafactual, você conclui que o programa reduziu as despesas médias com saúde em aproximadamente $14. Ao discutir um pouco mais os resultados com o consultor, você pergunta se as famílias que decidiram não se inscrever no programa não poderiam ser sistematicamente diferentes daquelas que se inscreveram. Por exemplo, as famílias que se registraram para o HISP podiam ser aquelas que esperavam ter maiores despesas com saúde, ou mesmo pessoas mais bem informadas sobre o programa ou pessoas mais preocupadas com a saúde de suas famílias. Alternativamente, talvez as famílias que se inscreveram fossem, em média,

Tabela 3.3 Caso 2 — Impacto do HISP Usando Inscritos e Não Inscritos (Comparação de Médias)

	Inscritos	Não inscritos	Diferença	Teste t
Despesas da família com a saúde	7,8	21,8	–13,9	–39,5

Fonte: Autores.

Tabela 3.4 Caso 2 — Impacto do HISP Usando Inscritos e Não Inscritos (Análise de Regressão)

	Regressão linear	Regressão linear multivariada
Impacto estimado nos gastos da família com saúde	–13,9** (0,35)	–9,4** (0,32)

Fonte: Autores.

Observação: Os erros padrão estão entre parênteses.

** 1% de significância estatística.

mais pobres do que as que não se inscreveram, visto que o HISP é voltado às famílias carentes. O consultor garante que a análise de regressão pode controlar potenciais diferenças entre os dois grupos. Ao controlar todas as características das famílias que estão disponíveis na base de dados, o consultor estima o impacto do programa, conforme apresentado na tabela 3.4.

Com uma regressão linear simples das despesas com saúde sobre uma variável categórica que indica se as famílias se inscreveram ou não no programa, você encontra um impacto estimado de menos $13,90; em outras palavras, você estima que o programa causou um decréscimo nas despesas médias em saúde de $13,90. No entanto, quando todas as outras características da população amostral são mantidas constantes, você estima que o programa reduziu as despesas das famílias inscritas em $9,40 ao ano.

PERGUNTA 2

A. Com base nos resultados do caso 2, o HISP deveria ser estendido nacionalmente?
B. Pode-se dizer que esta análise controla todos os fatores que determinam as diferenças nas despesas com saúde entre os dois grupos?

Notas

1. Observe que estamos assumindo ausência de atrição na amostra durante o período de dois anos, ou seja, nenhuma família terá deixado a amostra. Esta não é uma hipótese realista para a maioria das pesquisas envolvendo famílias. Na prática, famílias que mudam de residência às vezes não são encontradas e algumas famílias se desfazem e deixam de existir como unidade familiar.
2. Como outro exemplo, se os jovens que esperam se beneficiar do programa de capacitação têm maior probabilidade de se inscrever (porque esperam, por exemplo, salários mais altos com a capacitação), então estaremos comparando um grupo de indivíduos que esperavam rendas mais altas com um grupo de indivíduos que não esperam rendas mais altas.

CAPÍTULO 4

Métodos de Seleção Aleatória

Tendo discutido duas abordagens para a construção de cenários contrafatuais que são comumente usados, mas que têm um alto risco de viés — comparações *antes-e-depois* e *com-e-sem* — agora nos voltamos a um conjunto de métodos que podem ser aplicados para estimar com maior precisão os impactos de programas. Como veremos, contudo, obter tal estimativa nem sempre é tão simples como pode parecer à primeira vista. Os programas são, em sua maioria, concebidos e implementados em um ambiente complexo e mutável, em que muitos fatores podem influenciar os resultados, tanto para os participantes do programa quanto para aqueles que não participam.

Secas, terremotos, recessões, mudanças no governo e mudanças nas políticas locais e internacionais fazem parte do mundo real e, como avaliadores, queremos ter certeza de que o impacto estimado do nosso programa continua válido, apesar dessa miríade de fatores.

Como veremos ao longo desta parte do livro, as regras de inscrição dos participantes em um programa serão o parâmetro-chave para selecionar o método de avaliação de impacto. Acreditamos que, na maioria dos casos, os métodos de avaliação devem tentar se enquadrar no contexto de regras operacionais de um programa (com alguns ajustes aqui e ali) e não o contrário. Contudo, também partimos da premissa de que *todos os programas sociais devem ter regras justas e transparentes para a alocação do programa*. Uma das regras mais justas e mais transparentes de alocação de recursos escassos entre populações igualmente merecedoras acaba sendo dar a todos os elegíveis uma oportunidade

igual de participar do programa. Uma maneira de fazer isso é simplesmente realizar um sorteio. Neste capítulo, vamos examinar vários *métodos de seleção aleatória*, que são semelhantes a sorteios e que decidem quem entra em um programa em determinado momento e quem não entra. Esses métodos de seleção aleatória não apenas proporcionam aos administradores dos programas uma regra justa e transparente para a alocação de recursos escassos *entre populações igualmente merecedoras,* mas também representam os métodos mais robustos para avaliar o impacto de um programa.

Os métodos de seleção aleatória podem, frequentemente, derivar das regras de funcionamento de um programa. Para muitos programas, a população dos participantes pretendidos — isto é, o conjunto de todas as unidades que o programa gostaria de atender — é maior do que o número de participantes que de fato o programa pode acolher num determinado momento. Por exemplo, em um único ano, um programa de educação pode fornecer material escolar e um currículo atualizado a 500 dentre milhares de escolas elegíveis no país. Ou um programa de emprego para jovens pode ter uma meta de atingir 2.000 jovens desempregados no primeiro ano de operação, embora haja dezenas de milhares de jovens desempregados que o programa gostaria de atender. Por diversas razões, os programas podem ser incapazes de atingir toda a população de interesse. Restrições orçamentárias podem simplesmente impedir os administradores de oferecer o programa a todas as unidades elegíveis desde o princípio. Mesmo que exista o orçamento para cobrir um número ilimitado de participantes, algumas vezes restrições de capacidade impedirão que um programa se estenda a todos ao mesmo tempo. No exemplo do programa de capacitação para a juventude, o número de jovens desempregados que querem capacitação profissional pode ser maior do que o número de vagas disponíveis nas escolas técnicas durante o primeiro ano do programa, o que pode limitar o número de pessoas que podem se inscrever.

Na realidade, a maioria dos programas tem restrições orçamentárias ou operacionais que os impedem de alcançar todos os participantes pretendidos ao mesmo momento. Neste contexto, em que a população de participantes elegíveis é maior do que o número de vagas disponíveis no programa, os administradores do programa devem definir um mecanismo de racionamento para alocar os serviços do programa. Em outras palavras, alguém terá de tomar uma decisão sobre quem entrará no programa e quem não entrará. O programa poderia ser oferecido por ordem de chegada, ou com base em características observadas (por exemplo, primeiro mulheres e crianças, ou primeiro os municípios mais pobres); ou a seleção poderia ser baseada em características não observadas (por exemplo, deixando que indivíduos se inscrevam por sua própria vontade e conhecimento) ou, até mesmo, por um sorteio.

Alocação Aleatória do Tratamento

Quando um programa é alocado aleatoriamente a uma grande população elegível, podemos gerar uma estimativa robusta do cenário contrafactual, considerado o padrão ouro da avaliação de impacto. A alocação aleatória de tratamento basicamente usa um sorteio para decidir quem, entre as populações igualmente elegíveis, receberá o programa e quem não receberá.[1] Cada unidade elegível de tratamento (por exemplo, um indivíduo, domicílio, comunidade, escola, hospital ou outro) tem uma probabilidade igual de seleção para o tratamento.[2]

Antes de discutirmos como implementar a alocação aleatória na prática e por que isso gera um forte cenário contrafactual, consideremos a razão pela qual a alocação aleatória é também uma maneira justa e transparente de conceder serviços de programas. Uma vez que tenha sido definida uma população-alvo (por exemplo, domicílios abaixo da linha da pobreza, crianças com menos de 5 anos de idade ou escolas em áreas rurais), a alocação aleatória é uma regra de atribuição justa porque permite que os gestores do programa garantam que todas as pessoas ou unidades elegíveis tenham a mesma chance de receber o programa e que o programa não seja selecionado usando critérios arbitrários ou subjetivos, ou mesmo através de apadrinhamento ou outras práticas injustas. Quando existe excesso de demanda por um programa, a alocação aleatória é um método que pode ser facilmente explicado pelos gestores do programa e facilmente compreendido pela população. Quando o processo de seleção é realizado através de um processo aberto e replicável, o método de alocação aleatória não pode ser facilmente manipulado e, portanto, protege os gestores do programa de potenciais acusações de favoritismo ou corrupção. A alocação aleatória, portanto, tem seus próprios méritos como mecanismo de racionamento, que vão muito além de sua utilidade como ferramenta de avaliação de impacto. De fato, temos nos deparado com uma série de programas que usam rotineiramente sorteios como forma de selecionar os participantes do conjunto de indivíduos elegíveis, principalmente devido às suas vantagens para a administração e a governança.[3]

Por que a Alocação Aleatória Produz uma Excelente Estimativa do contrafactual?

Como discutido anteriormente, o grupo de comparação ideal será o mais semelhante possível ao grupo de tratamento em todos os aspectos, exceto com respeito à inscrição no programa que está sendo avaliado. O importante é que, quando selecionamos unidades aleatoriamente para designá-las aos

grupos de tratamento e de comparação, o processo de alocação aleatória em si produzirá dois grupos que têm uma alta probabilidade de serem estatisticamente iguais, desde que o número de participantes potenciais aos quais aplicamos o processo de alocação aleatória seja suficientemente grande. Especificamente, com um número suficientemente grande de observações, o processo de alocação aleatória produzirá grupos que têm médias estatisticamente equivalentes *para todas as suas características*. Por sua vez, essas médias também tendem para a média da população da qual são sorteadas.[4] A figura 4.1 ilustra por que a seleção aleatória produz um grupo de comparação estatisticamente equivalente ao grupo de tratamento. Suponha que a população de unidades elegíveis (participantes potenciais) consista de 1.000 pessoas, das quais a metade seja escolhida aleatoriamente e selecionada para o grupo de tratamento e a outra metade para o grupo de comparação. Poderíamos, por exemplo, escrever os nomes de todas as 1.000 pessoas em pedaços de papel, misturá-los num recipiente e, em seguida, pedir para alguém retirar 500 nomes às cegas. Se houver sido determinado que os primeiros 500 nomes constituam o grupo de tratamento, então se terá um grupo de tratamento selecionado aleatoriamente (os primeiros 500 nomes retirados) e um grupo de comparação selecionado aleatoriamente (os 500 nomes deixados no recipiente).

Agora suponhamos que, das 1.000 pessoas originais, 40 por cento sejam mulheres. Como os nomes foram selecionados aleatoriamente, dos 500 nomes retirados do recipiente, aproximadamente 40 por cento também serão mulheres. Se entre as 1.000 pessoas, 20 por cento tiverem olhos azuis, então aproximadamente 20 por cento - tanto do grupo de tratamento quanto do grupo de comparação - devem ter olhos azuis também. Em geral,

Figura 4.1 Características dos Grupos sob Alocação Aleatória do Tratamento

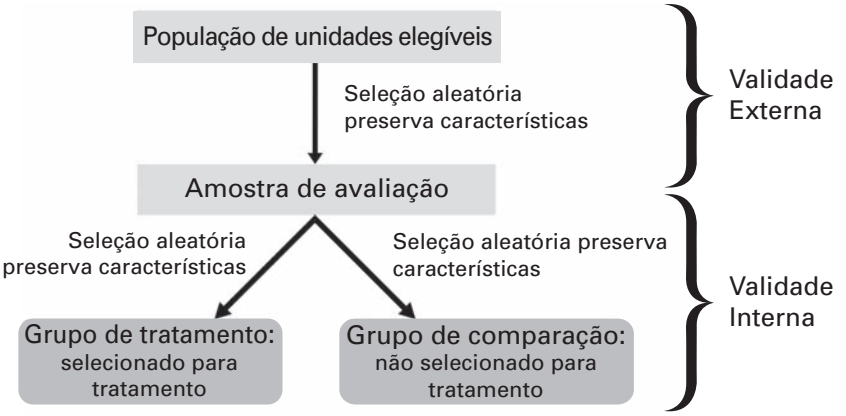
Fonte: Autores.

se a população de unidades elegíveis for suficientemente grande, então qualquer característica da população também será observada em ambos o grupo de tratamento e o de comparação. Podemos imaginar que, se características observadas - como o sexo ou a cor dos olhos de uma pessoa - se refletem tanto no grupo de tratamento quanto no de comparação, então, logicamente, características que são mais difíceis de observar (variáveis não observadas), tais como motivação, preferências ou outros traços de personalidade difíceis de medir também se refletirão em ambos os grupos. Assim, grupos de tratamento e de comparação gerados através de alocação aleatória serão semelhantes não apenas em suas características observadas, mas também nas suas características não observadas. Por exemplo, você pode não ser capaz de observar ou medir quão "simpáticas" são as pessoas, mas você sabe que, se 20 por cento das pessoas na população de unidades elegíveis são simpáticas, então aproximadamente 20 por cento das pessoas no grupo de tratamento serão simpáticas; o mesmo será verdade para o grupo de comparação. A alocação aleatória ajudará a garantir que, em média, os grupos de tratamento e de comparação sejam semelhantes em todos os sentidos, tanto nas características observadas como nas não observadas.

Quando uma avaliação utiliza a alocação aleatória para os grupos de tratamento e comparação, sabemos que, teoricamente, o processo deve produzir dois grupos equivalentes. Com os dados de linha de base na nossa amostra de avaliação, podemos testar esta hipótese empiricamente e verificar que, de fato, não há diferenças sistemáticas nas características observadas entre os grupos de tratamento e de comparação antes que o programa inicie. Então, depois de lançarmos o programa, se observarmos diferenças de resultados entre os grupos de tratamento e de comparação, saberemos que essas diferenças podem ser explicadas apenas pela introdução do programa, uma vez que, em sua composição, os dois grupos eram idênticos na linha de base e foram expostos aos mesmos fatores ambientais externos ao longo do tempo. Nesse sentido, o grupo de comparação *controla* todos os fatores que poderiam explicar também o resultado de interesse. Podemos ter bastante confiança de que a nossa estimativa de impacto médio, dada como a diferença entre o resultado sob o tratamento (o resultado médio do grupo de tratamento selecionado aleatoriamente) e a nossa estimativa do contrafactual (o resultado médio do grupo de comparação selecionado aleatoriamente) constituem o verdadeiro impacto do programa, uma vez que, por construção, eliminamos todos os fatores, observados e não observados, que poderiam plausivelmente explicar a diferença nos resultados.

Na figura 4.1 presume-se que todas as unidades da população elegível seriam alocadas ao grupo de tratamento ou de comparação. Em alguns casos, no entanto, não é necessária a inclusão de todos eles na avaliação.

Figura 4.2 Amostragem Aleatória e Alocação Aleatória do Tratamento



Fonte: Autores.

Por exemplo, se a população de unidades elegíveis inclui um milhão de mães e se pretende avaliar a eficácia de uma transferência monetária sobre a probabilidade de elas vacinarem seus filhos, pode ser suficiente uma amostra representativa de, digamos, mil mães e alocar essas mil ao grupo de tratamento ou ao de comparação. A figura 4.2 ilustra esse processo. Pela mesma lógica explicada acima, tomar uma amostra aleatória de unidades elegíveis da população para formar a amostra de avaliação preserva as características da população de unidades elegíveis. Mais uma vez, a seleção aleatória dos grupos de tratamento e de comparação a partir da amostra de avaliação preserva as características.

Validade externa e interna

Os passos destacados acima para a alocação aleatória de tratamento assegurarão tanto a validade interna quanto a externa da avaliação de impacto, desde que a amostra de avaliação seja suficientemente grande (figura 4.2).

Conceito-chave:
Uma avaliação é internamente válida se ela usar um grupo de comparação válido.

Validade interna significa que o impacto estimado do programa é livre de todos os outros potenciais fatores de perturbação, ou que o grupo de comparação representa o verdadeiro contrafactual, de maneira que estamos estimando o verdadeiro impacto do programa. Lembre-se que a alocação aleatória produz um grupo de comparação que é estatisticamente equivalente ao grupo de tratamento na linha de base antes do início do programa. Uma vez iniciado o programa, o grupo de comparação é exposto ao mesmo

conjunto de fatores externos ao longo do tempo, sendo o programa a única exceção. Portanto, se quaisquer diferenças nos resultados aparecerem entre os grupos de tratamento e de comparação, estas só podem ser devidas à existência do programa no grupo de tratamento. Em outras palavras, a validade interna de uma avaliação de impacto é garantida pelo processo de *alocação aleatória de tratamento*.

Validade externa significa que o impacto estimado na amostra de avaliação pode ser generalizado à população de todas as unidades elegíveis. Para que isto seja possível, a amostra deve ser representativa da população de unidades elegíveis. Na prática, isso significa que a amostra de avaliação deverá ser selecionada dentre a população, usando uma das muitas variações da *amostragem aleatória*.[5] Observe que levantamos dois tipos de alocação aleatória: uma com o propósito de amostragem (para validade externa) e uma como um método deavaliação de impacto (para validade interna). Uma avaliação de impacto pode gerar estimativas de impacto internamente válidas através de alocação aleatória de tratamento; no entanto, se a avaliação for realizada em uma amostra não aleatória da população, os impactos estimados podem não ser generalizáveis à população de unidades elegíveis. De forma inversa, se a avaliação utilizar uma amostra aleatória da população de unidades elegíveis mas o tratamento não for selecionado de forma aleatória, então a amostra será representativa, mas o grupo de comparação poderá não ser válido.

Conceito-chave:
Uma avaliação é externamente válida se a amostra representar precisamente a população de unidades elegíveis. Os resultados são, então, generalizáveis para a população de unidades elegíveis.

Quando a Alocação Aleatória pode ser Usada?

Na prática, a alocação aleatória deve ser considerada sempre que um programa tiver um excesso de demanda, isto é, quando o número de potenciais participantes for maior do que o número de vagas disponíveis no momento em que o programa tiver que ser implementado. Algumas circunstâncias também merecem alocação aleatória como ferramenta de avaliação, mesmo que os recursos do programa não sejam limitados. Por exemplo, governos podem querer usar alocação aleatória para testar programas novos ou potencialmente caros, com consequências intencionais e não intencionais desconhecidas. Nesse contexto, a alocação aleatória é justificada durante um período de avaliação piloto para testar rigorosamente os efeitos do programa antes de estendê-lo a uma população maior.

Existem duas circunstâncias nas quais a alocação aleatória é viável como método de avaliação de impacto:

1. *Quando a população elegível for maior do que o número de vagas disponíveis no programa*. Quando a demanda por um programa excede a oferta,

um sorteio simples pode ser usado para selecionar o grupo de tratamento dentro da população elegível. Nesse contexto, cada unidade da população tem uma chance igual de ser selecionada para o programa. O grupo que vence no sorteio é o grupo de tratamento e o restante da população, a quem não é oferecido o programa, forma o grupo de comparação. Enquanto existir uma restrição de recursos que impeça a ampliação do programa a toda a população, os grupos de comparação podem ser mantidos para medir os impactos de curto, médio e longo prazo do programa. Nesse contexto, nenhum dilema ético surge do fato de se manter um grupo de comparação indefinidamente, uma vez que uma parcela da população necessariamente será deixada de fora do programa.

Como exemplo, suponha que o Ministério da Educação queira oferecer bibliotecas escolares para escolas públicas de todo o país, mas o Ministério da Fazenda disponibiliza orçamento apenas para cobrir um terço delas. Se o Ministério da Educação quiser que cada escola pública tenha a mesma chance de receber uma biblioteca, ele realizará um sorteio no qual cada escola terá a mesma chance (1 em 3) de ser selecionada. As escolas sorteadas receberão uma nova biblioteca e constituirão o grupo de tratamento; aos dois terços restantes de escolas públicas do país não será oferecida a biblioteca e servirão de grupo de comparação. A menos que fundos adicionais sejam alocados ao programa de bibliotecas, permanecerá um grupo de escolas que não terão recursos para bibliotecas por meio do programa, podendo ser utilizadas como grupo de comparação para medir o contrafactual.

2. *Quando um programa tiver de ser implementado gradualmente até cobrir toda a população elegível.* Quando um programa é implementado gradualmente, a randomização da ordem em que os participantes recebem o programa dá a cada unidade elegível a mesma chance de receber tratamento na primeira fase ou em uma fase posterior do programa. Enquanto o "último" grupo ainda não tiver sido inserido no programa, servirá como grupo de comparação válido, a partir do qual se poderá estimar o contrafactual para os grupos já inseridos.

   Por exemplo, suponha que o Ministério da Saúde queira treinar todos os 15.000 enfermeiros no país a utilizar um novo protocolo de saúde, mas precise de três anos para treiná-los todos. No contexto de uma avaliação de impacto, o Ministério poderia escolher aleatoriamente um terço dos enfermeiros para receber treinamento no primeiro ano, um terço para receber treinamento no segundo ano e um terço para receber treinamento no terceiro ano. Para avaliar o efeito do programa de treinamento

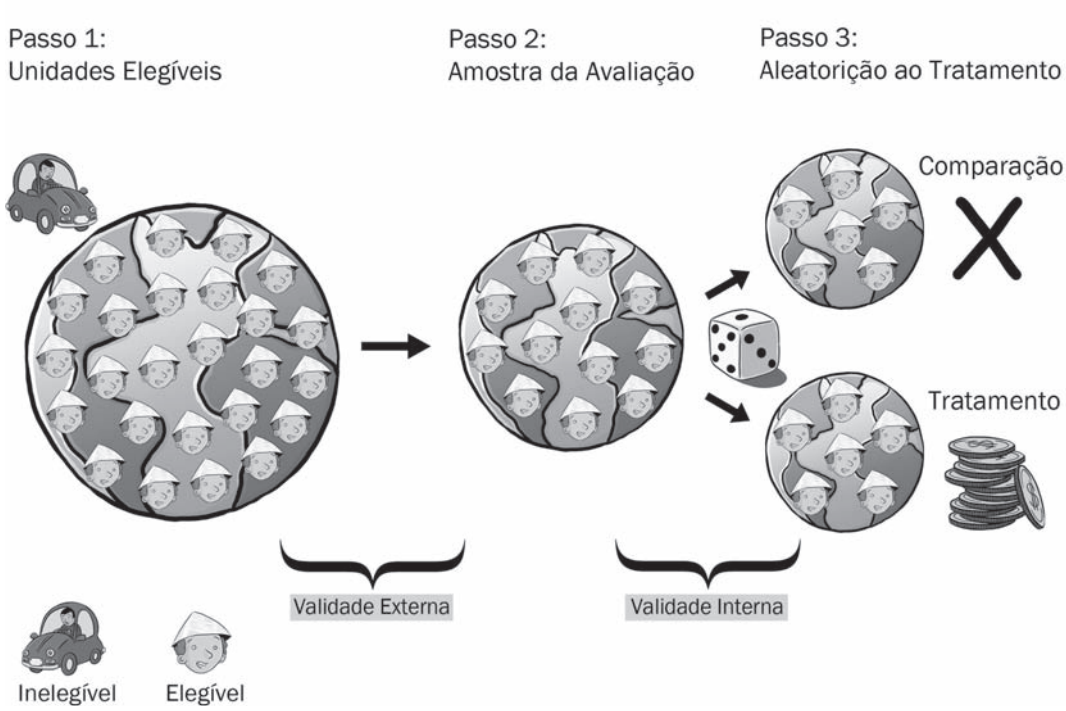
um ano após a sua implementação, o grupo de enfermeiros formados no primeiro ano constituiria o grupo de tratamento e o grupo de enfermeiros aleatoriamente selecionados para a formação no terceiro ano seria o grupo de comparação, uma vez que eles ainda não teriam recebido o treinamento.

Como se Aloca Aleatoriamente o Tratamento?

Agora que discutimos o que a alocação aleatória é capaz de fazer e por que ela produz um bom grupo de comparação, vamos tratar dos passos necessários para que uma alocação aleatória de tratamento seja bem sucedida. A figura 4.3 ilustra esse processo.

O primeiro passo na alocação aleatória é definir as unidades que são elegíveis para o programa. Dependendo do programa, uma unidade pode ser uma pessoa, um centro de saúde, uma escola, ou até mesmo um município inteiro.

Figura 4.3 Passos na Alocação Aleatória do Tratamento



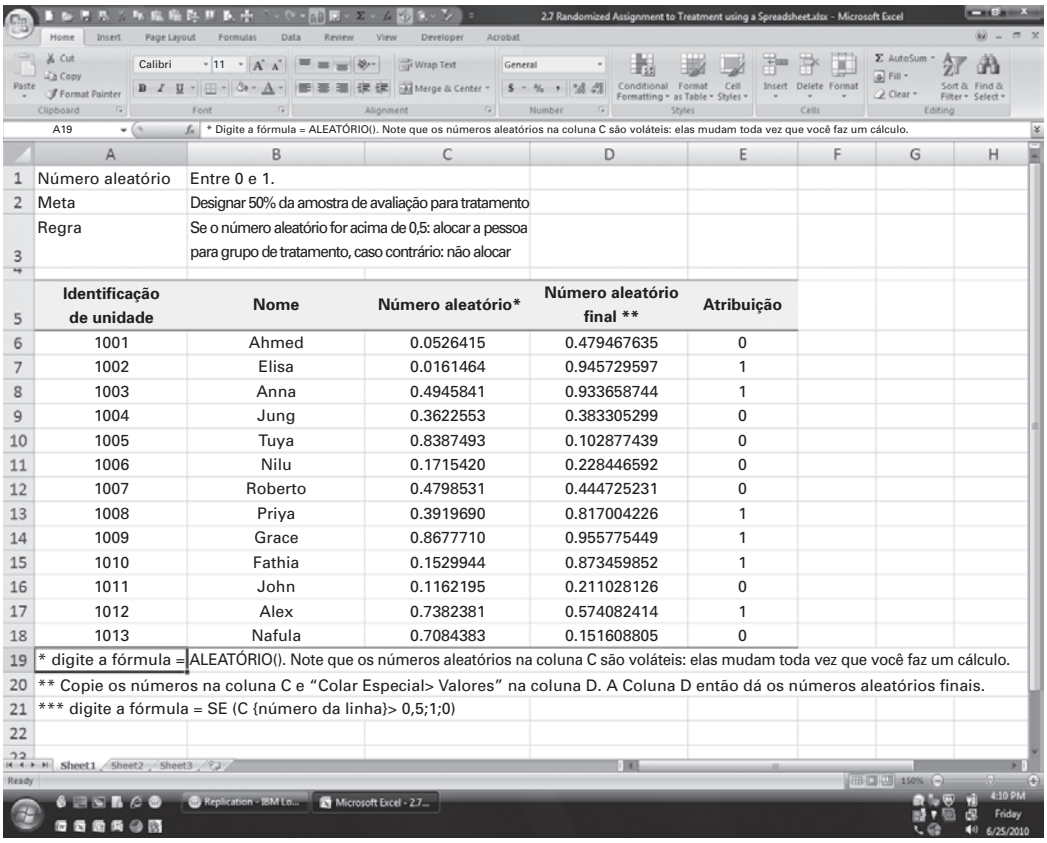
Fonte: Autores.

A população de unidades elegíveis consiste daquelas sobre as quais você está interessado em conhecer o impacto do programa. Por exemplo, se você está implementando um programa de capacitação de professores primários em áreas rurais, então professores de escolas secundárias ou professores de escolas primárias em áreas urbanas não pertenceriam a sua população de unidades elegíveis.

Depois de ter determinado a população das unidades elegíveis, será necessária a comparação do tamanho do grupo com o número de observações necessárias para a avaliação. Esse número é determinado através de cálculos de poder estatístico e depende dos tipos de perguntas que você gostaria de ver respondidas (ver capítulo 11). Se a população elegível for pequena, todas as unidades elegíveis podem ter de ser incluídas na avaliação. Alternativamente, se houver mais unidades elegíveis do que o necessário para a avaliação, o passo 2 consiste em selecionar uma amostra destas unidades da população para serem incluídas na amostra de avaliação. Note que esta segunda etapa é realizada principalmente para limitar os custos de coleta de dados. Se for verificado que dados dos sistemas de monitoramento existentes podem ser utilizados para a avaliação e que estes sistemas cobrem a população de unidades elegíveis, então você não precisará tirar uma amostra de avaliação separadamente. No entanto, imagine uma avaliação em que a população de unidades elegíveis inclua dezenas de milhares de professores em todas as escolas públicas do país e você precise coletar informações detalhadas sobre o conhecimento pedagógico dos professores. Entrevistar todo e cada professor pode não ser factível na prática, mas você pode descobrir que é suficiente tirar uma amostra de 1.000 professores, distribuídos por 100 escolas. Contanto que a amostra de escolas e professores seja representativa de toda a população de professores de escolas públicas, qualquer resultado encontrado na avaliação poderá ser generalizado para o resto dos professores e das escolas públicas no país. Coletar dados sobre essa amostra de 1.000 professores será, naturalmente, muito mais barato do que coletar dados sobre todos os professores em todas as escolas públicas do país.

Finalmente, no passo 3 serão formados os grupos de tratamento e de comparação, a partir das unidades da amostra de avaliação. Isso requer que você primeiro decida sobre uma regra de como designar os participantes com base em números aleatórios. Por exemplo, se você precisa designar 40 de 100 unidades da amostra de avaliação ao grupo de tratamento, você pode decidir designar estas 40 unidades com os mais altos números aleatórios ao grupo de tratamento e o restante ao grupo de comparação. Você, então, atribui um número aleatório a cada unidade de observação na amostra de avaliação, usando uma planilha ou software de estatística especializado (figura 4.4), e usa a regra previamente escolhida para formar os

Figura 4.4 Alocação Aleatória do Tratamento Usando uma Planilha



	A	B	C	D	E	F	G	H
1	Número aleatório	Entre 0 e 1.						
2	Meta	Designar 50% da amostra de avaliação para tratamento						
3	Regra	Se o número aleatório for acima de 0,5: alocar a pessoa para grupo de tratamento, caso contrário: não alocar						
4								
5	**Identificação de unidade**	**Nome**	**Número aleatório***	**Número aleatório final ****	**Atribuição**			
6	1001	Ahmed	0.0526415	0.479467635	0			
7	1002	Elisa	0.0161464	0.945729597	1			
8	1003	Anna	0.4945841	0.933658744	1			
9	1004	Jung	0.3622553	0.383305299	0			
10	1005	Tuya	0.8387493	0.102877439	0			
11	1006	Nilu	0.1715420	0.228446592	0			
12	1007	Roberto	0.4798531	0.444725231	0			
13	1008	Priya	0.3919690	0.817004226	1			
14	1009	Grace	0.8677710	0.955775449	1			
15	1010	Fathia	0.1529944	0.873459852	1			
16	1011	John	0.1162195	0.211028126	0			
17	1012	Alex	0.7382381	0.574082414	1			
18	1013	Nafula	0.7084383	0.151608805	0			
19	* digite a fórmula = ALEATÓRIO(). Note que os números aleatórios na coluna C são voláteis: elas mudam toda vez que você faz um cálculo.							
20	** Copie os números na coluna C e "Colar Especial> Valores" na coluna D. A Coluna D então dá os números aleatórios finais.							
21	*** digite a fórmula = SE (C {número da linha}> 0,5;1;0)							
22								

Fonte: Autores.

grupos de tratamento e de comparação. Note que é importante decidir sobre a regra antes de executar o software que dá às unidades seus números aleatórios; caso contrário, você pode ser tentado a decidir sobre uma regra com base nos números aleatórios que você vê e isso invalidaria a alocação aleatória.

A lógica por trás do processo automatizado não é diferente da alocação aleatória baseada em *cara ou coroa* ou em tirar nomes de um chapéu: é um mecanismo que determina aleatoriamente se cada unidade está no grupo de tratamento ou no grupo de comparação. Nos casos em que a alocação aleatória precisa ser feita em um fórum público, algumas técnicas mais "artesanais" de alocação aleatória podem ser utilizadas. Os exemplos a seguir presumem que a unidade de randomização seja uma pessoa:

1. Se você quiser alocar 50 por cento dos indivíduos ao grupo de tratamento e 50 por cento ao grupo de comparação, jogue a moeda para cada pessoa. Você tem que decidir antecipadamente se a *cara* ou a *coroa* na moeda atribuirá uma pessoa ao grupo de tratamento.

2. Se você quiser alocar um terço da amostra de avaliação ao grupo de tratamento, você pode lançar dados para cada pessoa. Primeiro, você tem que decidir uma regra. Por exemplo: ao cair, se o dado exibir o número 1 ou 2, isso poderia significar alocação no grupo de tratamento, ao passo que 3, 4, 5, ou 6 significaria uma atribuição ao grupo de comparação. Você lançaria o dado uma vez para cada pessoa na amostra de avaliação e as alocaria com base no número que aparecer.

3. Escreva os nomes de todas as pessoas em papéis do mesmo tamanho e forma. Dobre os papéis para que os nomes não possam ser vistos e misture-os completamente num chapéu ou algum outro recipiente. Antes de começar a retirar os papéis, decida as regras - ou seja, quantos papéis você irá retirar e se um nome retirado significa que estará sendo selecionado para o grupo de tratamento. Uma vez que a regra esteja clara, pergunte a alguém no meio da multidão (alguém imparcial - uma criança, por exemplo) para tirar tantos pedaços de papel quantos você precisar de participantes no grupo de tratamento.

Quer você use um sorteio público, quer lançamento de dados ou números aleatórios gerados por computador, é importante documentar o processo para garantir sua transparência. Isso significa, em primeiro lugar, que a regra de atribuição tem que ser decidida e comunicada a todos os membros do público. Segundo, você deve respeitar a regra ao retirar os números aleatórios; e, terceiro, você deve ser capaz de mostrar que o processo foi realmente aleatório. Nos casos de sorteios e lançamento de dados, você poderia filmar o processo; atribuições de números aleatórios baseadas em computador requerem que você forneça um registro computacional, de maneira que o processo possa ser replicado por auditores.[6]

Em que Nível Você Realiza a Alocação Aleatória?

A alocação aleatória pode ser feita em nível individual, familiar, comunitário, ou regional. Em geral, o nível em que atribuímos aleatoriamente as unidades aos grupos de tratamento e de comparação será muito afetado por onde e como o programa está sendo implementado. Por exemplo, se um programa de saúde estiver sendo implementado no nível dos postos de saúde, você deve primeiro escolher uma amostra aleatória de postos de saúde

e depois designar aleatoriamente alguns deles para o grupo de tratamento e outros para o grupo de comparação.

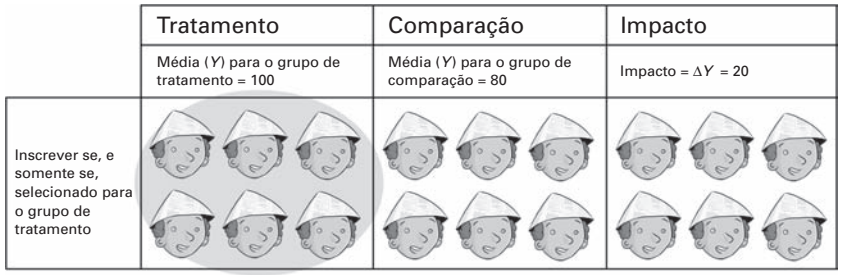
Quando o nível da alocação aleatória for mais elevado - por exemplo, em nível regional ou provincial em um determinado país - pode ser difícil realizar uma avaliação de impacto, porque o número de regiões ou províncias, na maioria dos países, não é suficientemente grande para produzir grupos de tratamento e de comparação balanceados. Por exemplo, se um país tem apenas seis províncias, isso permitiria apenas três províncias de tratamento e três de comparação, números que são insuficientes para assegurar que as características dos grupos de tratamento e de comparação sejam balanceadas.

Mas à medida que o nível de alocação aleatória se reduz, por exemplo, ao nível do indivíduo ou da família, as chances de efeitos de transbordamento e de contaminação aumentam.[7] Por exemplo, se o programa consiste em fornecer medicamentos vermífugos a domicílios e um domicílio do grupo de tratamento estiver localizado perto de um domicílio do grupo de comparação, o domicílio de comparação pode ser afetado positivamente por influência do tratamento no domicílio tratado, pois as chances de contraírem vermes dos vizinhos serão reduzidas. Os domicílios de tratamento e de comparação precisam estar localizados suficientemente distantes um do outro para evitar efeitos de transbordamento. Porém, à medida que aumenta a distância entre os domicílios, tornam-se mais dispendiosas tanto a implementação do programa quanto a realização de pesquisas. Como regra geral, se efeitos de transbordamento puderem ser razoavelmente descartados, o melhor é realizar a alocação aleatória do tratamento no menor nível possível de implementação do programa; isso irá assegurar que o número de unidades, tanto no grupo de tratamento quanto no de comparação, seja o maior possível. Os efeitos de transbordamento serão discutidos no capítulo 8.

Estimando o Impacto sob Alocação Aleatória

Uma vez tomada uma amostra aleatória de avaliação e alocado o tratamento de forma aleatória, é bastante fácil estimar o impacto do programa. Após o programa operar por algum tempo, os resultados terão que ser medidos tanto para as unidades de tratamento quanto as de comparação. O impacto do programa é simplesmente a diferença entre o resultado médio (Y) para o grupo de tratamento e o resultado médio (Y) para o grupo de comparação. Por exemplo, na figura 4.5, o resultado médio para o grupo de tratamento é igual a 100 e o resultado médio para o grupo de comparação é igual a 80, de maneira que o impacto do programa é igual a 20.

Figura 4.5 Estimando o Impacto sob Alocação Aleatória



Fonte: Autores.

Estimando o impacto do Programa de Subsídio ao Seguro Saúde sob Alocação Aleatória

Vamos voltar agora ao exemplo do Programa de Subsídio ao Seguro Saúde (HISP) e verificar o que significa "alocação aleatória" neste contexto. Lembre-se que você está tentando estimar o impacto do programa a partir de um piloto que envolve 100 municípios de tratamento.

Tendo realizado duas avaliações de impacto usando contrafatuais potencialmente enviesados (e tendo obtido recomendações conflitantes de políticas - ver capítulo 3), você decide voltar à prancheta de desenho para repensar a forma de obter um contrafactual mais preciso. Após novas deliberações com a sua equipe de avaliação, você está convencido de que construir uma estimativa válida do contrafactual exigirá a identificação de um grupo de municípios que sejam idênticos aos 100 municípios de tratamento em todos os aspectos, com a única exceção de que um grupo participou do HISP e o outro, não. Como o HISP foi lançado como piloto e os 100 municípios de tratamento foram selecionados aleatoriamente dentre todos os municípios rurais do país, você observa que os municípios deveriam, em média, ter as mesmas características que a população geral dos municípios rurais. O contrafactual pode, portanto, ser estimado de forma válida medindo-se os gastos com saúde dos domicílios elegíveis nos municípios que não fizeram parte do programa.

Por sorte, no momento das pesquisas de linha de base e de seguimento, a empresa de pesquisas coletou dados em outros 100 municípios rurais onde o programa não havia sido oferecido na primeira rodada. Estes 100 municípios adicionais também foram escolhidos aleatoriamente do conjunto de municípios elegíveis, o que significa que eles também terão, em média, as mesmas características que a população geral dos municípios rurais.

Assim, a maneira como os dois grupos de municípios foram escolhidos assegura que eles tenham características idênticas, exceto pelo fato de que os 100 municípios de tratamento receberam o HISP e os 100 municípios de comparação, não. Ocorreu uma alocação aleatória do tratamento.

Dada a alocação aleatória do tratamento, você está bastante confiante de que não há fatores externos, com exceção da HISP, que poderiam explicar as diferenças de resultados entre os municípios de tratamento e de comparação. Para validar esta hipótese, você testa se os domicílios elegíveis nos municípios de tratamento e de comparação têm características semelhantes na linha de base, como mostra a tabela 4.1.

Você observa que as características médias dos domicílios nos municípios de tratamento e de comparação são, de fato, muito similares. A única diferença estatisticamente significativa é o número de anos de educação do cônjuge e essa diferença é pequena. Note que, mesmo com um experimento aleatório numa grande amostra, pode-se encontrar um pequeno número de diferenças.[8] Com a validade do grupo de comparação estabelecida, a sua

Tabela 4.1 Caso 3 — Balanço entre os Municípios de Tratamento e de Comparação na Linha de Base

Características do Domicílio	Municípios de Tratamento (N = 2964)	Municípios de Comparação (N = 2664)	Diferença	teste *t*
Despesas com saúde ($ anual per capita)	14,48	14,57	–0,09	–0,39
Idade do chefe da família (anos)	41,6	42,3	–0,7	–1,2
Idade do cônjuge (anos)	36,8	36,8	0,0	0,38
Educação do chefe da família (anos)	2,9	2,8	0,1	2,16*
Educação do cônjuge (anos)	2,7	2,6	0,1	0,006
Chefe da família é do sexo feminino = 1	0,07	0,07	–0,0	–0,66
Indígena = 1	0,42	0,42	0,0	0,21
Número de membros do domicílio	5,7	5,7	0,0	1,21
Tem banheiro = 1	0,57	0,56	0,01	1,04
Hectares de terra	1,67	1,71	–0,04	–1,35
Distância do hospital (km)	109	106	3	1,02

Fonte: Cálculos do autor.

* Significativo a 5%.

estimativa do contrafactual agora é a média dos gastos com saúde dos domicílios elegíveis nos 100 municípios de comparação (Tabela 4.2).

Considerando que agora você tem uma estimativa válida do contrafactual, você pode determinar o impacto da HISP simplesmente tomando a diferença entre os gastos com saúde nos domicílios elegíveis dos municípios de tratamento e a estimativa do contrafactual. O impacto é uma redução de $10,10 ao longo de dois anos. Replicar essa estimativa com uso da análise de regressão produz o mesmo resultado, como mostra o tabela 4.3.

Com a alocação aleatória, podemos ter certeza de que não há fatores sistematicamente diferentes entre os grupos de tratamento e de comparação e que também possam explicar a diferença de gastos com saúde. Ambos os conjuntos de municípios foram expostos ao mesmo conjunto de políticas e programas nacionais durante os dois anos de tratamento. Assim, a razão mais plausível para que os domicílios pobres em comunidades de tratamento tenham gastos menores do que os domicílios dos municípios de comparação é que o primeiro grupo recebeu o programa de seguro de saúde e o outro grupo, não.

Tabela 4.2 Caso 3 — Impacto do HISP Usando a Alocação Aleatória (Comparação de Médias)

	Tratamento	Comparação	Diferença	teste *t*
Despesas das famílias com saúde (na linha de base)	14,48	14,57	–0,09	–0,39
Despesas das famílias com saúde (no seguimento)	7,8	17,9	–10,1**	–25,6

Fonte: Cálculos do autor.

** Significativo a 1%.

Tabela 4.3 Caso 3 — Impacto do HISP Usando a Alocação Aleatória (Análise de Regressão)

	Regressão Linear	Regressão Linear Multivariada
Impacto estimado nas despesas das famílias com saúde	–10,1**	–10,0**
	(0,39)	(0,34)

Fonte: Cálculos do autor.

Nota: Erros padrão estão entre parênteses.

** Significativo a 1 %.

QUESTÃO 3

A. Por que a estimativa de impacto decorrente da regressão linear multivariada não é alterada, considerando-se que outros fatores são controlados?

B. Com base no impacto estimado no caso 3, o HISP deveria ser ampliado nacionalmente?

Alocação Aleatória em Ação

A alocação aleatória é frequentemente usada em trabalhos rigorosos de avaliação de impacto, tanto em avaliações de larga escala quanto nas menores. A avaliação do programa Progresa, no México (Schultz 2004), é uma das avaliações de grande escala mais conhecidas que faz uso da alocação aleatória (quadro 4.1).

Quadro 4.1: Transferências Condicionadas de Renda e a Educação no México

O programa Progresa, agora chamado de "Oportunidades", começou em 1998 e oferece transferências de renda para as mães pobres no México rural, condicionadas à matrícula de seus filhos na escola, com presença confirmada pelo professor. Este programa social de larga escala foi um dos primeiros a ser concebido com uma avaliação rigorosa em mente e a alocação aleatória foi usada para ajudar a identificar o efeito das transferências condicionadas de renda em diversos resultados, em especial na matrícula escolar.

As bolsas, para crianças da 3ª à 9ª série, somam cerca de 50 a 75 por cento do custo privado com educação e são garantidas por três anos. As comunidades e famílias elegíveis para o programa foram determinadas com base em um índice de pobreza criado a partir de dados do censo e da coleta de dados de linha de base. Devido à necessidade de se implementar este programa social de larga escala gradativamente, cerca de dois terços das localidades (314 de 495) foram selecionadas aleatoriamente para receber o programa nos dois primeiros anos; as 181 restantes serviram como grupo de controle, antes de entrarem no programa no terceiro ano.

Com base na alocação aleatória, Schultz (2004) constatou um aumento médio no número de matrículas de 3,4 por cento para todos os alunos da 1ª à 8ª série, com o maior aumento observado entre as meninas que haviam concluído a 6 ª série - 14,8 por cento. A razão provável é que as meninas tendem a abandonar a escola em maior proporção à medida que crescem e, por isso, receberam uma transferência um pouco maior para que permaneçam na escola após o ensino primário. Estes impactos de curto prazo foram, então, extrapolados para prever o impacto de longo prazo do Progresa na escolaridade ao longo da vida e nos ganhos de renda.

Fonte: Schultz 2004.

a. Para ser exato, Schultz combinou alocação aleatória com métodos de diferenças-em-diferenças. O capítulo 8 discute os benefícios de se combinar diversas metodologias de avaliação de impacto.

Duas Variações de Alocação Aleatória

Consideremos agora duas variantes que se valem das propriedades da alocação aleatória: a oferta aleatória de tratamento e a promoção aleatória de tratamento.

Oferta Aleatória: quando nem Todos Cumprem com sua Alocação

Anteriormente, na discussão sobre alocação aleatória, consideramos que o administrador do programa tem o poder de alocar as unidades a grupos de tratamento e de comparação, com os encaminhados para o tratamento recebendo o programa e aqueles selecionados para o grupo de comparação sem recebê-lo. Em outras palavras, assumimos que as unidades que foram designadas aos grupos de tratamento e de comparação *cumpriram com* sua alocação. O pleno cumprimento da alocação é mais frequentemente atingido em condições de laboratório ou em experimentos médicos, nos quais o pesquisador pode se assegurar de que, primeiro, todos os indivíduos do grupo de tratamento tomam o medicamento e, segundo, que nenhum dos indivíduos do grupo de comparação o tome.[9]

Nos programas sociais da vida real, o pleno cumprimento dos critérios de seleção do programa (e, portanto, a conformidade da adesão aos grupos de tratamento e de comparação) é um ideal perseguido e os gestores de políticas e os avaliadores de impacto igualmente se esforçam para chegar o mais próximo possível desse ideal. Na prática, no entanto, o cumprimento rigoroso de 100 por cento das designações de tratamento e de comparação, como planejado, pode não ocorrer, apesar dos melhores esforços do implementador do programa e do avaliador de impacto. Só porque um professor é selecionado para o grupo de tratamento e lhe é oferecido capacitação não significa que ele ou ela irá, realmente, aparecer ao primeiro dia do curso. Da mesma forma, um professor que é selecionado para o grupo de comparação pode acabar encontrando uma maneira de participar do curso. Sob tais circunstâncias, uma comparação direta do grupo originalmente selecionado para o tratamento com o grupo originalmente selecionado para a comparação produzirá uma estimativa da *"intenção-de-tratar" (ITT, do inglês 'intention-to-treat')*. Ou seja, estaremos comparando aqueles que pretendíamos tratar (os selecionados para o grupo de tratamento) com aqueles que não pretendíamos tratar (os selecionados para o grupo de comparação). Por si só, essa é uma medida de impacto muito interessante e relevante, uma vez que a maioria dos formuladores de políticas e administradores de programas podem apenas oferecer o programa, não podendo forçar que a população-alvo participe dele.

Por outro lado, podemos também ter interesse em estimar o impacto do programa sobre aqueles que realmente participam ou aceitam o tratamento. Fazê-lo requer correção para o fato de que algumas das unidades selecionadas ao grupo de tratamento, na verdade, não receberam o tratamento, ou que algumas das unidades selecionadas para o grupo de comparação, na verdade, o receberam. Em outras palavras, queremos estimar o impacto do programa sobre aqueles a quem o tratamento foi oferecido *e* que realmente participaram. Esta é a estimativa *"efeito do tratamento sobre os tratados"* (*TOT, do inglês 'treatment-on-the-treated'*).

Oferta Aleatória de um Programa e Participação Final

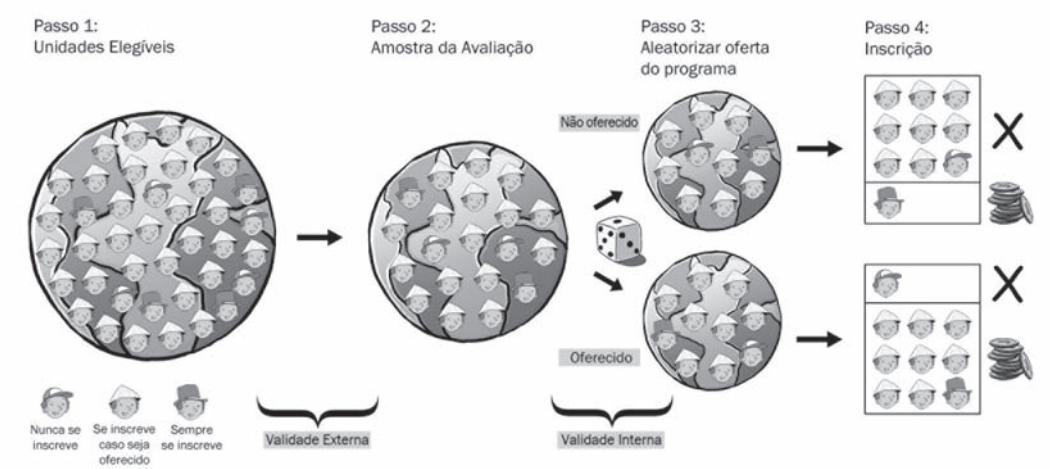
Imagine que você está avaliando o impacto de um programa de treinamento sobre os salários dos indivíduos. O programa é alocado aleatoriamente, em nível individual, e é oferecido ao grupo de tratamento, mas não ao grupo de comparação. Muito provavelmente, você encontrará três tipos de indivíduos na população:

- *Participa-se-oferecido*. Estes são os indivíduos que cumprem com a sua alocação. Se forem selecionados para o grupo de tratamento (em que o programa é oferecido), eles participam, ou se inscrevem; se forem selecionados para o grupo de comparação (em que o programa não é oferecido), eles não se inscrevem.
- *Nunca*. Estes são os indivíduos que nunca se inscrevem ou participam do programa, mesmo que sejam selecionados para o grupo de tratamento. Eles são os "desobedientes" do grupo de tratamento.
- *Sempre*. Estes são os indivíduos que vão encontrar uma maneira de se inscrever no programa ou participar dele, mesmo que sejam selecionados para o grupo de comparação. Eles são os "desobedientes" do grupo de comparação.

No contexto do programa de capacitação, o grupo *Nunca* pode ser de pessoas desmotivadas que, mesmo que lhes sejam oferecidas vagas no curso, não aparecem. No grupo *Sempre*, em contrapartida, as pessoas são tão motivadas que encontram uma maneira de entrar no programa, mesmo que tenham sido originalmente designadas para o grupo de comparação. No grupo *Participa-se-oferecido* estão aqueles que se matriculam no curso se este lhes for oferecido (o grupo de tratamento), mas não procuram se matricular se forem selecionados para o grupo de comparação.

A Figura 4.6 apresenta a oferta aleatória do programa e a matrícula final, ou participação, quando os grupos *Participa-se-oferecido, Nunca* e

Figura 4.6 Oferta Aleatória de um Programa



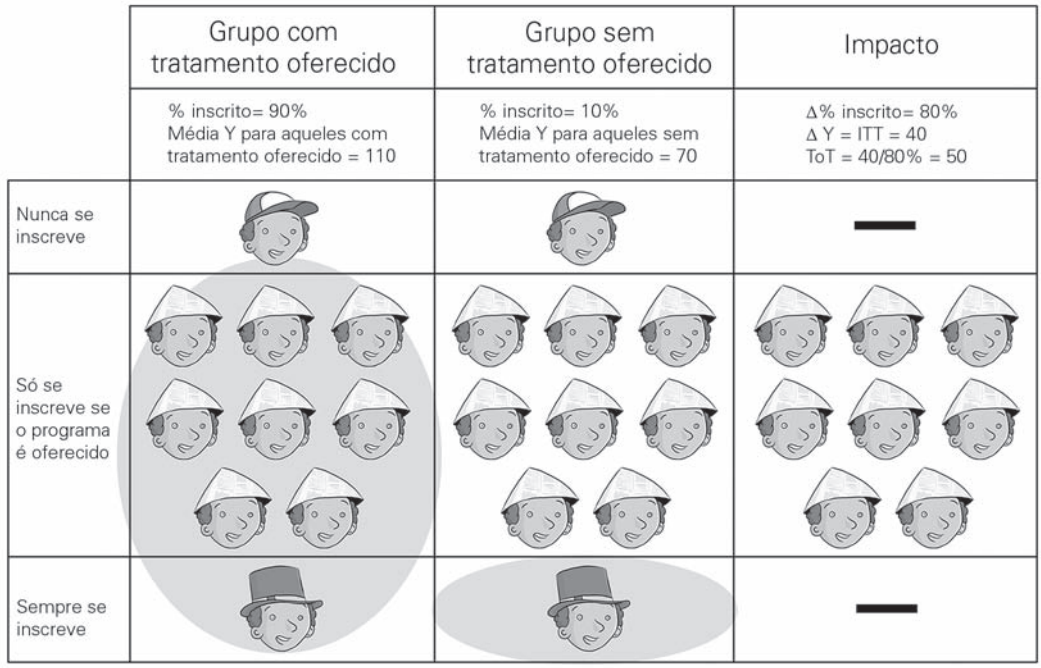
Fonte: Autores.

Sempre estão presentes. Nós assumimos que a população de unidades tenha 80 por cento de *Participa-se-oferecido*, 10 por cento de *Nunca* e 10 por cento de *Sempre*. Se tomarmos uma amostra aleatória da população na amostra de avaliação, a amostra de avaliação também terá cerca de 80 por cento de *Participa-se-oferecido,* 10 por cento de *Nunca* e 10 por cento de *Sempre*. Então, se dividirmos aleatoriamente a amostra de avaliação em um grupo de tratamento e um grupo de comparação, devemos voltar a ter cerca de 80 por cento de *Participa-se-oferecido,* 10 por cento de *Nunca* e 10 por cento de *Sempre* em ambos os grupos. No grupo ao qual é oferecido o tratamento, os indivíduos *Participa-se-oferecido* e *Sempre* se matricularão e apenas as pessoas *Nunca* ficarão de fora. No grupo ao qual não é oferecido o tratamento, os *Sempre* se inscreverão, enquanto os grupos *Participa-se-oferecido* e *Nunca* ficarão de fora.

Estimando o Impacto sobre a Oferta Aleatória

Tendo estabelecido a diferença entre a oferta de um programa e a efetiva inscrição ou participação, recorremos a uma técnica que pode ser utilizada para estimar o impacto do tratamento sobre os tratados - isto é, o impacto do programa sobre aqueles aos quais foi oferecido tratamento *e* que realmente se inscreveram. Essa estimativa é realizada em dois passos, ilustrados na Figura 4.7[10].

Figura 4.7 Estimando o Impacto do Tratamento nos Indivíduos Tratados no caso de Oferta Aleatória



Fonte: Autores.

Nota: ITT é a estimativa da "intenção-de-tratar" obtida comparando-se os resultados daqueles aos quais o tratamento foi oferecido com aqueles aos quais o tratamento não foi oferecido (independentemente da inscrição ocorrer). TOT é a estimativa do "efeito do tratamento sobre os tratados", ou seja, o impacto do programa estimado naqueles aos quais foi proposto tratamento e que efetivamente se inscreveram. As figuras sobre o fundo sombreado são aquelas que efetivamente se inscrevem.

Primeiro, estima-se o impacto da intenção de tratar. Lembre-se que esta é apenas a diferença direta no indicador de resultado (Y) entre o grupo ao qual é oferecido o tratamento e o grupo ao qual não é oferecido o tratamento. Por exemplo, se a renda média (Y) do grupo de tratamento for de \$110 e a renda média do grupo de comparação for de \$70, então a estimativa da *intenção-de-tratar* do impacto (ITT) será de \$40.

Segundo, precisamos recuperar a estimativa do *efeito-do-tratamento-sobre-os-tratados* (TOT), a partir da estimativa de *intenção-de-tratar*. Para tanto, precisamos identificar de onde veio a diferença de \$40. Vamos proceder por eliminação. Em primeiro lugar, sabemos que a diferença não pode ser causada por quaisquer diferenças entre os *Nunca* nos grupos de

tratamento e de comparação. A razão é que os *Nunca* jamais se inscrevem no programa, de maneira que, para eles, não faz diferença estarem no grupo de tratamento ou no grupo de comparação. Em segundo lugar, sabemos que a diferença de $40 não pode ser causada por diferenças entre as pessoas *Sempre* nos grupos de tratamento e de comparação, porque as pessoas *Sempre* se inscrevem no programa sempre. Para elas, também não faz diferença se estão no grupo de tratamento ou no grupo de comparação. Assim, a diferença nos resultados entre os dois grupos deve, necessariamente, vir do efeito do programa no único grupo afetado pela designação ao grupo de tratamento ou ao grupo de comparação, isto é, o grupo *Participa-se-oferecido*. Portanto, se pudermos identificar os *Participa-se-oferecido* em ambos os grupos, será fácil estimar o impacto do programa sobre eles.

Na realidade, embora saibamos que esses três tipos de indivíduos existem na população, não podemos separar univocamente os indivíduos entre *Participa-se-oferecido, Nunca* ou *Sempre*. No grupo ao qual foi oferecido tratamento, podemos identificar os *Nunca* (porque não se inscreveram), mas não podemos diferenciar os *Sempre* dos *Participa-se-oferecido* (porque ambos estão inscritos). No grupo ao qual não foi oferecido tratamento, podemos identificar o grupo *Sempre* (porque se inscrevem no programa), mas não podemos diferenciar entre os *Nunca* e os *Participa-se-oferecido.*

No entanto, uma vez que observemos que 90 por cento das unidades do grupo de tratamento se inscrevem, podemos deduzir que 10 por cento das unidades de nossa população devem ser de *Nunca* (que é a fração de indivíduos no grupo ao qual foi oferecido tratamento que não se inscreveram). Além disso, se observarmos que 10 por cento das unidades do grupo ao qual não foi oferecido tratamento se inscrevem, sabemos que 10 por cento são *Sempre* (novamente, a fração de indivíduos no grupo ao qual não foi oferecido tratamento que se inscreveram). Isso deixa 80 por cento das unidades no grupo *Participa-se-oferecido*. Sabemos que todo o impacto de $40 veio de uma diferença no número de inscrições para 80 por cento das unidades em nossa amostra que são *Participa-se-oferecido*. Agora, se 80 por cento das unidades são responsáveis por um impacto médio de $40 para todo o grupo ao qual é oferecido o tratamento, então o impacto sobre estes 80 por cento de *Participa-se-oferecido* deve ser 40/0,8, ou $50. Dito de outra forma, o impacto do programa para os *Participa-se-oferecidos* é de $50, mas quando este impacto está espalhado por todo o grupo ao qual é oferecido tratamento, o efeito médio é diluído pelos 20 por cento que não cumprem com a alocação aleatória original.

Lembre-se que um dos problemas básicos com a auto-seleção nos programas é que você não pode sempre saber por que algumas pessoas escolhem participar e outras não. Quando designamos aleatoriamente unidades

ao programa, mas a participação delas é voluntária, ou possa existir uma maneira de unidades do grupo de comparação entrarem no programa, então temos um problema semelhante: nós nem sempre compreenderemos os processos comportamentais que determinam se um indivíduo se comporta como um *Nunca*, um Sempre *ou um Participa*-se-*oferecido* no nosso exemplo acima. No entanto, desde que o não cumprimento não seja muito grande, a alocação aleatória inicial ainda fornece uma ferramenta poderosa para estimar o impacto. O lado negativo da atribuição aleatória com cumprimento imperfeito é que a estimativa de impacto já não é válida para toda a população. Em vez disso, aplica-se apenas a um subgrupo específico dentro da nossa população-alvo, os *Participa-se-oferecidos.*

A oferta aleatória de um programa tem duas características importantes que nos permitem estimar o impacto, mesmo sem o cumprimento integral dos indivíduos (ver quadro 4.2):[11]

1. Ela pode servir como um preditor da inscrição efetiva no programa, se a maioria das pessoas se comportar como um *Participa-se-oferecido*, inscrevendo-se no programa quando lhe for oferecido o tratamento ou não se inscrevendo quando não lhe for oferecido o tratamento.

2. Como os dois grupos (tratamento oferecido e não oferecido) são gerados por um processo de seleção aleatória, as características dos indivíduos nos dois grupos não estão correlacionadas com qualquer outra coisa, como capacidade ou motivação, que também possa afetar os resultados (*Y*).

Promoção Aleatória ou Desenho de Incentivo

Na seção anterior, vimos como estimar o impacto com base na alocação aleatória de tratamento, mesmo quando o cumprimento da alocação aos grupos de tratamento e de comparação originais é incompleto. A seguir, iremos propor uma abordagem muito semelhante, que pode ser aplicada para avaliar os programas que têm elegibilidade universal ou inscrições abertas ou nos quais o administrador do programa não possa controlar quem participa e não participa.

É comum os governos implementarem programas para os quais é difícil excluir potenciais participantes ou forçá-los a participar. Muitos programas permitem que os potenciais participantes optem por se inscrever e não são, portanto, capazes de excluir potenciais participantes que queiram se inscrever. Além disso, alguns programas têm um orçamento grande o suficiente para estender o programa a toda população elegível de uma só vez, de maneira que escolher aleatoriamente grupos de tratamento e de comparação excluindo potenciais participantes por causa da avaliação não

Quadro 4.2: Oferta Aleatória de *Vouchers* Escolares na Colômbia

O Programa de Ampliação da Cobertura da Educação Secundária (Programa de Ampliación de Cobertura de la Educación Secundaria [PACES]), na Colômbia, proporcionou *vouchers* a mais de 125.000 estudantes para cobrir um pouco mais da metade dos custos de frequentar uma escola secundária particular. Por causa do orçamento limitado do PACES, os *vouchers* foram alocados por sorteio. Angrist et al. (2002) aproveitaram esse tratamento selecionado aleatoriamente para determinar o efeito do programa de *vouchers* sobre os resultados educacionais e sociais.

Eles descobriram que os ganhadores do sorteio tinham 10 por cento a mais de probabilidade de completar a 8ª série e tiveram um desempenho, em média, 0,2 desvios-padrão superior em testes padronizados, três anos após o sorteio inicial. Eles também descobriram que os efeitos educacionais foram maiores no caso das meninas que no dos meninos. Os pesquisadores, então, se debruçaram sobre o impacto do programa sobre vários resultados não educacionais e descobriram que os vencedores do sorteio tinham menos probabilidade de serem casados e trabalhavam cerca de 1,2 horas a menos por semana.

Houve algum descumprimento do desenho aleatório, na medida em que cerca de 90 por cento dos contemplados pelo sorteio haviam, realmente, usado o *voucher* ou outra forma de bolsa de estudos e 24 por cento dos não contemplados pelo sorteio tinham, na verdade, recebido bolsas de estudos. Angrist e seus colegas, portanto, também utilizaram a *intenção-de-tratar*, ou o status de ganho ou perda no sorteio, como uma variável instrumental para o *efeito-do-tratamento-sobre-os-tratados*, ou o recebimento efetivo de bolsas. Finalmente, os pesquisadores conseguiram realizar uma análise de custo-benefício para melhor compreender o impacto do programa de *vouchers*, tanto nas despesas das famílias quanto nas do governo. Eles concluíram que os custos sociais totais do programa são pequenos e compensados pelos retornos esperados pelos participantes e suas famílias, sugerindo, assim, que programas do lado da demanda, como o PACES, podem ser uma maneira custo-eficiente para aumentar o nível educacional.

Fonte: Angrist et al. 2002.

seria ético. Precisamos, portanto, de uma forma alternativa para avaliar o impacto destes tipos de programas — aqueles com inscrição voluntária e aqueles com cobertura universal.

Programas de inscrição voluntária tipicamente permitem que indivíduos interessados no programa se inscrevam e participem por conta própria. Imagine novamente o programa de treinamento discutido anteriormente mas, desta vez, sem que a alocação aleatória seja possível, sendo que qualquer indivíduo que deseje se inscrever no programa possa fazê-lo. Em consonância com o nosso exemplo anterior, esperamos

encontrar três tipos de pessoas: os que seguem as regras do programa, um grupo do tipo *Nunca* e um grupo do tipo *Sempre*. Como no caso anterior, as pessoas do tipo *Sempre* sempre se inscreverem no programa e as pessoas do tipo *Nunca* nunca se inscreverem. Mas e os cumpridores das regras? Nesse contexto, qualquer indivíduo que queira se inscrever no programa é livre para fazê-lo. E o que dizer de indivíduos que podem estar muito interessados em se inscrever mas que, por várias razões, podem ter informações suficientes ou o incentivo correto para que se inscrevam? Os cumpridores das regras, neste contexto, serão precisamente este grupo. Eles são os que *participam se incentivados:* um grupo de indivíduos que só se inscrevem no programa se lhes for dado um incentivo adicional ou um incentivo que os motive a se inscrever. Sem esse estímulo adicional, os *participam-se-incentivados* simplesmente ficariam fora do programa.

Mais uma vez voltando ao exemplo do programa de capacitação, se a agência que organiza a capacitação é bem financiada e tem capacidade suficiente para treinar todos os que queriam ser treinados, então o programa de formação profissional poderá estar aberto a qualquer pessoa desempregada que queira participar. É improvável, contudo, que todos os desempregados realmente queiram participar ou mesmo saibam da existência do programa. Algumas pessoas desempregadas podem ser relutantes em se inscrever porque sabem muito pouco sobre o conteúdo da capacitação e acham que é difícil conseguir informações adicionais. Agora, vamos supor que a agência de formação profissional contrate um agente comunitário para percorrer toda a cidade e arregimentar pessoas desempregadas para o programa de capacitação para o trabalho. Carregando uma lista de pessoas desempregadas, ele bate à porta, descreve o programa de capacitação e se oferece para ajudar a pessoa a se inscrever no programa ali mesmo. Obviamente, ele não pode obrigar ninguém a participar. Além disso, os desempregados que o agente comunitário não visita também podem se inscrever, mas terão de ir à agência para fazê-lo. Assim, temos agora dois grupos de pessoas desempregadas — as que foram visitadas pelo agente comunitário e as que não foram. Se o esforço do agente for eficaz, a proporção de inscrições entre as pessoas visitadas deve ser maior que a proporção entre as pessoas desempregadas que não foram visitadas.

Agora vamos pensar sobre como podemos avaliar este programa de capacitação profissional. Como sabemos, não podemos simplesmente comparar os desempregados que se inscrevem àqueles que não se inscrevem. A razão é que os desempregados que se inscrevem são, provavelmente, muito diferentes dos que não se inscrevem, tanto com respeito a variáveis observáveis quanto às não observáveis: eles podem ser mais

instruídos (isto pode ser observado facilmente), e eles estão, provavelmente, mais motivados e ansiosos para encontrar um emprego (o que é difícil de observar e medir).

No entanto, temos alguma variação adicional que podemos explorar para encontrar um grupo de comparação válido. Vamos considerar, por um momento, se podemos comparar o grupo que recebeu a visita do agente comunitário ao grupo que não foi visitado. Ambos os grupos contêm pessoas muito motivadas *(Sempre)*, que se inscreverão quer o agente comunitário bata à sua porta, quer não. Ambos os grupos também contêm pessoas desmotivadas *(Nunca)*, que não se inscreverão no programa apesar dos esforços do agente comunitário. E, finalmente, algumas pessoas *(Participa-se-incentivado)* vão se inscrever no treinamento se o agente comunitário as visitar, mas não vão se inscrever caso o agente não vier a procurá-las.

Se o agente comunitário tiver selecionado as pessoas aleatoriamente em sua lista de visitação, nós teríamos condições de usar o *efeito-do-tratamento-sobre-os-tratados* discutido anteriormente. A única diferença seria que, em vez de *oferecer* o programa aleatoriamente, o estaríamos *promovendo* aleatoriamente. Contanto que pessoas de perfil *Participa-se-incentivado* (que se inscrevem quando nós chegamos a elas, mas não se inscrevem se não chegarmos a elas) apareçam, teremos uma variação entre o grupo *com* a promoção ou divulgação e o grupo *sem* a promoção ou divulgação que nos permitirá identificar o impacto da capacitação no grupo *Participa-se-incentivado*. Em vez de responder à oferta do tratamento, os *Participam-se-incentivados* agora estão respondendo à promoção.

Queremos que a estratégia de divulgação seja eficaz e aumente substancialmente a inscrição do grupo *Participa-se-incentivado*. Ao mesmo tempo, não queremos que as atividades de promoção sejam tão amplas e eficazes que influenciem o resultado de interesse. Por exemplo, se os agentes comunitários tivessem oferecido grandes quantidades de dinheiro para levar os desempregados a se inscrever, seria difícil determinar se as alterações posteriores na renda foram causadas pela formação ou pela própria divulgação ou promoção.

A promoção aleatória é uma estratégia criativa que gera o equivalente a um grupo de comparação para fins de avaliação de impacto. Ela pode ser usada quando é possível organizar uma ação de promoção destinada a uma amostra aleatória da população de interesse. Leitores com algum conhecimento de econometria podem reconhecer novamente a terminologia apresentada na seção anterior: a promoção aleatória é uma variável instrumental que nos permite criar variação entre as unidades e explorar tal variação para criar um grupo de comparação válido.

Você Disse "Promoção"?

A promoção aleatória procura aumentar a participação em um programa voluntário de uma subamostra da população. Ela pode tomar várias formas. Por exemplo, pode-se optar por iniciar uma campanha de informação para alcançar as pessoas que não se inscreveram porque não sabiam ou não compreenderam totalmente o conteúdo do programa. Como alternativa, pode-se optar por fornecer incentivos para inscrição, como a oferta de pequenos brindes ou prêmios ou, ainda, disponibilizar transporte ou outra forma de auxílio.

Uma série de condições devem ser atendidas para a metodologia de promoção aleatória produzir uma avaliação de impacto válida.

Conceito-chave:
Promoção aleatória é um método semelhante à oferta aleatória. Em vez de selecionar unidades aleatoriamente a quem oferecemos o tratamento, selecionamos aleatoriamente as unidades às quais promoveremos o tratamento. Desse modo, podemos deixar o programa aberto a todos.

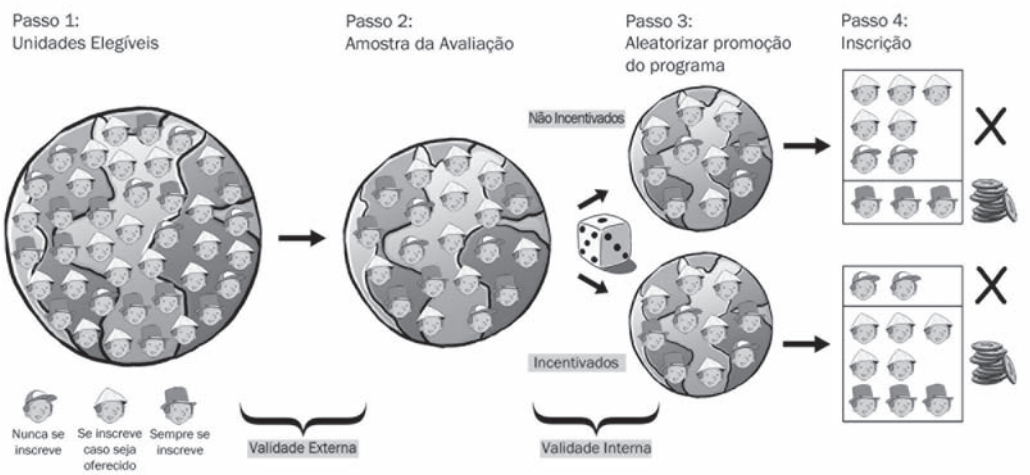
1. Os grupos promovidos e não promovidos devem ser comparáveis. As características dos dois grupos devem ser semelhantes. Isto é alcançado ao se atribuir aleatoriamente as atividades de divulgação ou de promoção às unidades da amostra de avaliação.
2. A campanha de divulgação deve aumentar a inscrição das pessoas do grupo promovido muito mais do que a das pessoas do grupo não promovido. Isso pode ser confirmado verificando-se se a taxa de inscrição é mais alta no grupo que recebe a promoção do que no grupo que não a recebe.
3. É importante que a promoção em si não afete diretamente o resultado de interesse, de modo que possamos dizer que as mudanças nos resultados de interesse são causadas pelo próprio programa e não pela promoção.

O Processo de Promoção Aleatória

O processo de promoção aleatória é apresentado na figura 4.8. Como nos métodos anteriores, começamos com a população de unidades elegíveis para o programa. Em contraste com a alocação aleatória, não podemos mais escolher aleatoriamente quem vai receber o programa e quem não vai receber o programa porque o programa é totalmente voluntário. No entanto, dentro da população de unidades elegíveis, haverá três tipos de unidades:

- *Sempre* — aqueles que sempre vão querer se inscrever no programa
- *Participa-se-incentivado* — aqueles que vão se inscrever no programa apenas se lhes for oferecida uma vantagem adicional
- *Nunca* — aqueles que nunca vão querer se inscrever no programa, independentemente se lhes ofereçamos alguma vantagem ou não.

Figura 4.8 Promoção Aleatória



Fonte: Autores.

Novamente, note que ser um *Sempre,* um *Participa-se-incentivado,* ou um *Nunca* é uma característica intrínseca das unidades que não pode ser medida pelo avaliador do programa, porque está relacionada a fatores como a motivação intrínseca e a inteligência.

Uma vez que a população elegível seja definida, o próximo passo é selecionar aleatoriamente uma amostra da população para fazer parte da avaliação. Estas são as unidades para as quais coletaremos dados. Em alguns casos — por exemplo, quando temos dados referentes a toda a população de unidades elegíveis — podemos decidir incluir toda a população na amostra de avaliação.

Uma vez que a amostra de avaliação seja definida, a promoção aleatória aloca aleatoriamente as unidades da amostra de avaliação entre o grupo promovido e o grupo não promovido. Uma vez que estamos escolhendo aleatoriamente os membros de ambos os grupos, estes compartilharão as características gerais da amostra de avaliação que, por sua vez, serão equivalentes às características da população de unidades elegíveis. Portanto, ogrupo promovido e o grupo não promovido terão características semelhantes.

Depois do fim da campanha de promoção, poderemos observar as taxas de inscrição nos grupos promovidos e não promovidos. No grupo não promovido, somente os *Sempre* se inscreverão. Embora, desta forma, possamos identificar as unidades *Sempre* no grupo não promovido, não teremos

condições de distinguir entre os *Nunca* e os *Participa-se-incentivado* neste grupo. Por outro lado, no grupo promovido, tanto os *Participa-se-incentivado* quanto os *Sempre* vão se inscrever, enquanto que os *Nunca* não se inscreverão. Assim, no grupo promovido seremos capazes de identificar o grupo *Nunca,* mas não conseguiremos distinguir entre os *Participa-se-incentivado* e os *Sempre*.

Estimando o Impacto sobre a Promoção Aleatória

Estimar o impacto de um programa usando promoção aleatória é um caso especial do método *efeito-do-tratamento-sobre-os-tratados* (figura 4.9). Imagine que a campanha de promoção eleve a inscrição de 30 por cento no grupo não promovido (3 indivíduos do tipo *Sempre*) para 80 por cento no grupo promovido (3 indivíduos do tipo *Sempre* e 5 do tipo *Participa-se-incentivado*). Suponha que o resultado médio de todos os indivíduos no grupo não promovido (10 indivíduos) seja 70, e que o resultado médio para todos os indivíduos do grupo promovido (10 indivíduos) seja 110. Qual seria, então, o impacto do programa?

Primeiro, podemos calcular a diferença direta entre o grupo promovido e o não promovido, que é de 40. Sabemos também que nenhuma parte dessa diferença de 40 provém dos *Nunca*, uma vez que eles não se inscrevem em

Figura 4.9 Estimando o Impacto sobre a Promoção Aleatória

Fonte: Autores.

Nota: As figuras sobre o fundo sombreado são aquelas que se inscrevem.

nenhum dos grupos. Do mesmo modo, sabemos que a diferença de 40 não pode vir dos *Sempre,* porque eles se inscrevem em ambos os grupos.

O segundo passo é recuperar o impacto que o programa teve nos *Participa-se-incentivados.* Sabemos que o efeito total médio de 40 pode ser atribuído aos *Participa-se-incentivados,* que constituem apenas 50 por cento da população. Para avaliar o efeito médio do programa em um indivíduo cumpridor das regras de participação, dividimos 40 pela porcentagem de *Participa-se-incentivados* na população. Embora não possamos identificar diretamente os *Participa-se-incentivados,* somos capazes de deduzir qual deve ser a sua *porcentagem* na população: a diferença entre as taxas de inscrição no grupo promovido e no não promovido (50 por cento, ou 0,5) Portanto, o efeito médio do programa no indivíduo que obedece as regras de participação é 40/0,5 = 80.

Considerando que a promoção é alocada aleatoriamente, os grupos promovido e não promovido têm características iguais, em média. Assim, as diferenças que observamos nos resultados médios entre os dois grupos devem ser causadas pelo fato de que no grupo promovido os *Participam-se-incentivados* se inscrevem, ao passo que no grupo não promovido eles não se inscrevem[12].

Usando a Promoção Aleatória para Estimar o Impacto do Programa de Subsídio ao Seguro Saúde

Vamos agora tentar usar o método de promoção aleatória para avaliar o impacto do HISP. Suponha que o Ministério da Saúde tome uma decisão executiva de que o subsídio ao seguro saúde deva estar disponível imediatamente a qualquer família que queira se inscrever no programa. No entanto, você sabe que, de fato, a expansão do programa nacionalmente será feita de forma gradativa ao longo do tempo. Assim, você faz um acordo para acelerar a inscrição no programa em um subconjunto aleatório de municípios, por meio de uma campanha de divulgação. Você realiza um intenso esforço de promoção em uma subamostra aleatória de municípios, incluindo campanhas de comunicação e marketing social destinadas a conscientização da população sobre o HISP. Após dois anos de promoção e implementação do programa, você descobre que 49,2 por cento dos domicílios em municípios que foram aleatoriamente selecionados para a promoção se inscreveram no programa, enquanto que apenas 8,4 por cento dos domicílios em municípios não promovidos se inscreveram (tabela 4.4).

Como os municípios promovidos e não promovidos foram selecionados aleatoriamente, você sabe que as características médias dos dois grupos devem ser as mesmas na ausência do programa. Você pode verificar esta

hipótese comparando as despesas de saúde de linha de base (bem como qualquer outra característica) das duas populações. Depois de dois anos de implementação do programa, você observa que a despesa média com saúde nos municípios promovidos é de $14,90, comparada a $18,80 em áreas não promovidas (uma diferença de menos $3,90). No entanto, como a única diferença entre os municípios promovidos e não promovidos é que os promovidos têm maior inscrição no programa (graças à promoção), esta diferença de 3,90 dólares em despesas de saúde deve ser devida aos 40,4 por cento de domicílios que se inscreveram nos municípios promovidos devido à promoção. Portanto, temos que ajustar a diferença em despesas de saúde para conseguir encontrar o impacto do programa nos *Participa-se-incentivado*. Para isso, dividimos a diferença entre os grupos promovidos pela porcentagem de *Participa-se-incentivado:* –3.9/0.404 = -$9,65. O seu colega, que teve aulas de econometria, estima o impacto do programa através do método dos mínimos quadrados em dois estágios (MQ2E) e encontra os resultados exibidos na tabela 4.5. Esse impacto estimado é válido para as famílias que só se inscreveram no programa devido à promoção - em outras

Tabela 4.4 Caso 4 — Impacto do HISP Usando a Promoção Aleatória (Comparação de Médias)

	Municípios Promovidos	Municípios não Promovidos	Diferença	teste *t*
Despesas das famílias com saúde (na linha de base)	17,1	17,2	–0,1	–0,47
Despesas das famílias com saúde (no seguimento)	14,9	18,8	–3,9	–18,3
Inscrição em HISP	49,2%	8,4%	40,4%	

Fonte: Cálculos do autor.

** Significativo a 1%.

Tabela 4.5 Caso 4 — Impacto do HISP Usando Promoção Aleatória (Análise de Regressão)

	Regressão Linear	Regressão Linear Multivariada
Impacto estimado sobre as despesas das famílias com saúde	–9,4** (0,51)	–9,7** (0,45)

Fonte: Cálculos do autor.

Nota: Desvios padrão estão entre parênteses.

** Significativo a 1%.

palavras, é o impacto sobre os *Participa-se-incentivado*. Para extrapolar esse resultado para toda a população, temos de assumir que todos os outros domicílios teriam reagido de forma semelhante, caso tivessem se inscrito no programa.

QUESTÃO 4

A. Quais são os pressupostos básicos necessários para aceitar o resultado do caso 4?
B. Com base no resultado do caso 4, o HISP deveria ser ampliado nacionalmente?

A Promoção Aleatória em Ação

O método de promoção aleatória pode ser utilizado em vários contextos. Gertler, Martinez e Vivo (2008) usaram o método para avaliar um programa de seguro saúde maternal e infantil na Argentina. Após a crise econômica de 2001, o governo da Argentina observou que os indicadores de saúde da população começaram a se deteriorar e, em particular, que a mortalidade infantil estava aumentando. Decidiu introduzir um regime nacional de seguro para as mães e seus filhos, que deveria ser ampliado para todo o país dentro de um ano. Ainda assim, funcionários do governo queriam avaliar o impacto do programa para ter certeza de que estava realmente melhorando o estado de saúde da população. Como poderia ser encontrado um grupo de comparação se cada mãe e criança no país tivesse o direito de se inscrever no regime de seguro, se assim o desejasse? Os dados das primeiras províncias que implementaram a intervenção mostraram que apenas 40 a 50 por cento dos domicílios estavam efetivamente se inscrevendo no programa. Assim, o governo lançou uma campanha de divulgação intensiva procurando informar às famílias sobre o programa. Contudo, a campanha de divulgação foi implementada apenas em uma amostra aleatória de municípios e não em todo o país.

Outros exemplos incluem a assistência de organizações não governamentais em uma avaliação da gestão escolar comunitária no Nepal e o Fundo de Investimento Social Boliviano (detalhados no quadro 4.3).

Limitações do Método de Promoção Aleatória

A promoção aleatória é uma estratégia útil para avaliar o impacto de programas voluntários e programas com elegibilidade universal, particularmente porque não exige a exclusão de qualquer unidade elegível. No entanto, essa abordagem tem algumas limitações dignas de nota, em comparação à alocação aleatória do tratamento.

Em primeiro lugar, a estratégia de promoção deve ser eficaz. Se a campanha de promoção não aumentar a inscrição, não aparecerá qualquer

Quadro 4.3: Promovendo Investimentos de Infraestrutura Educacional na Bolívia

Em 1991, a Bolívia institucionalizou e ampliou um exitoso Fundo de Investimento Social (SIF, do inglês *Social Investment Fund*), que fornecia financiamento às comunidades rurais para realizar pequenos investimentos em educação, saúde e infraestrutura de água. O Banco Mundial, que ajudava a financiar o SIF, conseguiu incluir uma avaliação de impacto no desenho do programa.

Como parte da avaliação de impacto do componente educação, as comunidades na região do Chaco foram selecionadas aleatoriamente para promoção ativa da intervenção do SIF e receberam visitas adicionais de motivação para se inscreverem no programa. O programa foi aberto a todas as comunidades elegíveis na região e foi baseado na demanda, de forma que as comunidades tinham de se candidatar aos fundos para um projeto específico.

Nem todas as comunidades se inscreveram no programa, mas a aceitação foi maior entre as comunidades promovidas.

Newman et al. (2002) utilizaram a promoção aleatória como uma variável instrumental. Eles concluíram que os investimentos em educação conseguiram melhorar medidas de qualidade da infraestrutura escolar, como energia elétrica, saneamento, livros por aluno e a proporção aluno-professor. No entanto, eles detectaram pouco impacto sobre os resultados educacionais, exceto por uma diminuição de cerca de 2,5 por cento na taxa de abandono escolar. Como resultado destes resultados, o Ministério da Educação e o SIF agora concentram mais atenção e recursos na parte pedagógica da educação, financiando melhorias de infraestrutura física somente quando fizerem parte de uma intervenção integrada.

Fonte: Newman et al. 2002.

diferença entre o grupo promovido e o grupo não promovido e, portanto, não haverá nada a comparar. Assim, é crucial realizar um piloto substancial da campanha de promoção, para ter-se certeza de que ela será efetiva. O lado positivo é que o desenho da campanha de promoção pode ajudar os gestores do programa, ensinando-lhes como aumentar as inscrições.

Em segundo lugar, a metodologia estima o impacto do programa apenas para um subconjunto da população de unidades elegíveis. Especificamente, o impacto médio do programa é calculado a partir do grupo de pessoas que se inscrevem no programa somente quando estimuladas a fazê-lo. No entanto, indivíduos desse grupo podem ter características muito diferentes dos indivíduos que sempre ou nunca se inscrevem e, portanto, o efeito médio do tratamento para toda a população pode ser diferente do efeito do tratamento médio estimado para os indivíduos que participam apenas quando estimulados.

Notas

1. Alocação aleatória de tratamento também é comumente referida como "experimentos aleatórios controlados", "avaliações aleatórias", "avaliações experimentais," e "experimentos sociais", entre outros termos.
2. Note que esta probabilidade não significa, necessariamente, uma chance 50-50 de vencer o sorteio. De fato, a maioria das avaliações de alocação aleatória atribuirá uma probabilidade de seleção determinada a cada unidade elegível, de maneira que o número de vencedores (tratamentos) seja igual ao número total de benefícios disponíveis. Por exemplo, se um programa tem verba suficiente para atender apenas 1.000 comunidades, de uma população de 10.000 comunidades elegíveis, então a cada comunidade será dada uma chance de 1 em 10 de ser selecionada para o tratamento. O poder estatístico (um conceito discutido em mais detalhe no capítulo 11) será maximizado quando a amostra de avaliação for dividida igualmente entre os grupos de tratamento e controle. Neste exemplo, para um tamanho de amostra total de 2.000 comunidades, o poder estatístico será maximizado quando for tirada uma subamostra de 1.000 comunidades de tratamento e uma subamostra de 1.000 comunidades de controle, em vez de tomar-se uma amostra aleatória simples de 20 por cento das 10.000 comunidades elegíveis originais (o que produziria uma amostra de avaliação de cerca de 200 comunidades de tratamento e 1.800 comunidades de controle).
3. Por exemplo, programas de habitação que oferecem casas subsidiadas usam sorteios rotineiramente para selecionar os participantes do programa.
4. Essa propriedade vem da Lei dos Grandes Números.
5. Uma amostra de avaliação pode ser estratificada por subtipos da população e também pode ser agrupada por unidades de amostragem. O tamanho da amostra dependerá do tipo particular de amostragem utilizado (ver parte 3).
6. A maioria dos programas de software permite que você defina um "número semente" para tornar os resultados da alocação aleatória totalmente transparentes e replicáveis.
7. Discutiremos conceitos como *transbordamento* ou *contaminação* em mais detalhes no capítulo 8.
8. Por razões estatísticas, nem todas as características observadas têm de ser semelhantes nos grupos de tratamento e de comparação para a aleatoriedade ser bem sucedida. Como regra geral, a escolha aleatória será considerada bem sucedida se cerca de 95 por cento das características observadas forem semelhantes. Por "semelhante", queremos dizer que não podemos rejeitar a hipótese nula de que as médias são diferentes entre os dois grupos quando se utiliza um intervalo de confiança de 95 por cento. Mesmo quando as características dos dois grupos forem verdadeiramente iguais, pode-se esperar que cerca de 5 por cento das características vai aparecer com uma diferença estatisticamente significativa.
9. Note-se que em ciências médicas, os pacientes no grupo de comparação tipicamente recebem um placebo, isto é, algo como uma pílula de açúcar que não deve ter nenhum efeito sobre o resultado pretendido. Isso é feito para

adicionalmente se controlar o "efeito placebo", ou seja, as possíveis mudanças de comportamento e resultados de se receber um tratamento, mesmo que o tratamento em si seja ineficaz.

10. Estas duas etapas correspondem à técnica econométrica de mínimos-quadrados-em-dois-estágios, o que produz um efeito de tratamento médio local.
11. Leitores com um conhecimento em econometria podem reconhecer o conceito: em termos estatísticos, a oferta aleatória do programa é usada como uma variável instrumental para a inscrição real. As duas características apresentadas são exatamente o que seria necessário de uma boa variável instrumental:
    - A variável instrumental deve ser correlacionada com a participação no programa.
    - A variável instrumental não pode ser correlacionada com os resultados (Y) (exceto através da participação do programa) ou com variáveis não observadas.
12. Mais uma vez, os leitores familiarizados com econometria podem reconhecer que o impacto é estimado pelo uso de "alocação aleatória aos grupos promovidos e não promovidos" como uma variável instrumental para o efetivo registro no programa.

Referências

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M.(2002). Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment. *American Economic Review* 92 (5): 1535–58.

Gertler, P., Martinez, S. & Vivo, S. (2008). Child-Mother Provincial Investment Project *Plan Nacer*. Washington, DC: Universidade da Califórnia em Berkeley e Banco Mundial.

Newman, J., Pradhan, M., Rawlings, L., Ridder, G., Coa, R., & Evia, J.L.. (2002). An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund. *World Bank Economic Review* 16 (2): 241–74.

Schultz, P. (2004). School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics* 74 (1): 199–250.

CAPÍTULO 5

Método de Regressão Descontínua

Os programas sociais geralmente usam um índice para decidir quem é e quem não é elegível a se inscrever no programa. Programas de combate à pobreza, por exemplo, focam normalmente em famílias pobres, identificadas por um escore ou índice de pobreza. O escore de pobreza pode se basear em uma fórmula aproximada dos recursos familiares, que mede o conjunto dos ativos básicos da família. As famílias com baixos escores são classificadas como pobres e as famílias com escores altos são consideradas relativamente prósperas. As autoridades do programa geralmente determinam um limiar ou nível de corte, abaixo do qual as famílias são consideradas pobres e elegíveis para o programa. Exemplos incluem o programa mexicano Progresa (Buddelmeyer e Skoufias 2004) e o sistema da Colômbia para selecionar beneficiários de investimentos sociais, o conhecido SISBEN (Barrera-Osorio, Linden e Urquiola 2007).

Os programas de aposentadoria são outro exemplo de um tipo de programa que focaliza as unidades de intervenção com base em um índice de elegibilidade contínuo, embora de tipo diferente. A idade constitui um índice contínuo e a idade de aposentadoria constitui o ponto de corte que determina a elegibilidade. Em outras palavras, somente as pessoas acima de certa idade são elegíveis para receber a pensão. Um terceiro exemplo de índice de elegibilidade contínuo poderia ser as pontuações em testes de aprendizagem. Muitos países concedem bolsas ou prêmios para os alunos que obtêm o melhor desempenho em testes padronizados, cujos resultados

são ranqueados do menor para o maior desempenho. Assim, se o número de bolsas for limitado, somente estudantes que obtiverem pontuação acima de uma nota de corte (tais como os melhores 15% do total de alunos) são elegíveis para a bolsa.

Conceito-chave:
O modelo de regressão descontínua (RDD) é adequado para programas que utilizam um índice contínuo para classificar potenciais participantes e que têm um valor de corte associado ao índice, que determina se os potenciais participantes serão beneficiados ou não pelo programa.

O modelo de regressão descontínua (RDD, *do inglês regression discontinuity design*) é um método de avaliação de impacto que pode ser usado em programas que apresentem um índice de elegibilidade contínuo, com um escore de corte claramente definido para determinar quem é elegível e quem não é. São necessárias duas condições para que o modelo de regressão descontínua seja aplicado.

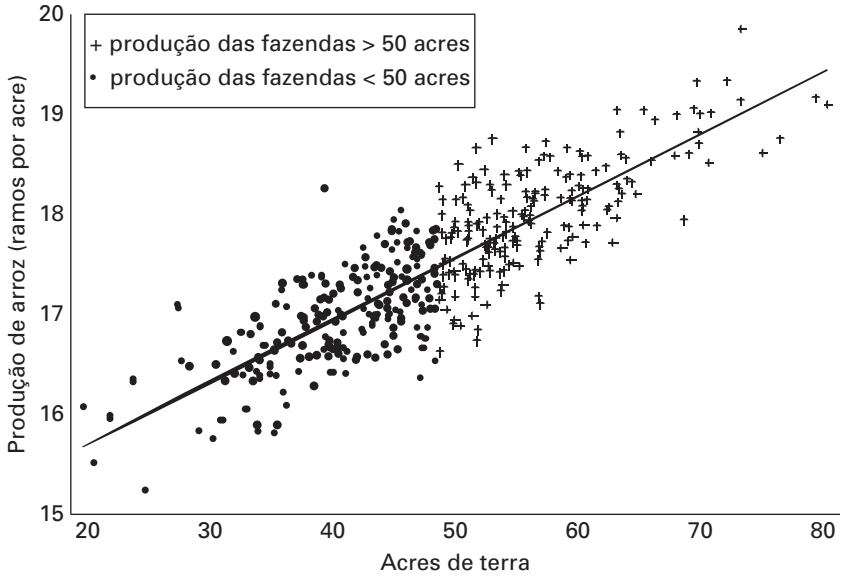
1. Um índice de elegibilidade contínuo - em outras palavras, uma medida contínua a partir da qual a população de interesse possa ser ranqueada, como um índice de pobreza, uma nota em teste de aprendizagem ou a idade.

2. Um escore de corte claramente definido - isto é, um ponto no índice acima ou abaixo do qual a população se classifica como elegível para o programa. Por exemplo, famílias com um valor de índice de pobreza inferior a 50 de um total de 100 poderiam ser classificadas como pobres; indivíduos com 67 anos ou mais de idade poderiam ser classificados como aposentados e estudantes com 90 pontos ou mais (de um total de 100) em um teste de aprendizagem poderiam ser elegíveis para uma bolsa de estudo. Os pontos de corte nestes exemplos são 50, 67 e 90, respectivamente.

Caso 1: Subsídios para Fertilizantes na Produção de Arroz

Considere um programa agrícola que subsidia a compra de fertilizantes por produtores de arroz, com o objetivo de melhorar a produção. O programa visa pequenas e médias fazendas, classificadas como fazendas com menos de 50 acres de área total. Antes do início do programa, podemos esperar que a relação entre o tamanho da fazenda e o total da produção de arroz seja conforme apresentado na figura 5.1, de modo que as fazendas menores têm uma produção menor do que as fazendas maiores. O escore de elegibilidade, neste caso, é o total de acres da fazenda e o ponto de corte é 50 acres. Segundo as regras de elegibilidade do programa, as fazendas abaixo de 50 acres são elegíveis para receber o subsídio de fertilizantes e as fazendas com 50 acres ou mais não são. Neste caso, podemos esperar algumas fazendas com 48, 49 ou 49,9 acres que participam do programa.

Figura 5.1 Produção de Arroz



Fonte: Autores.

Outro grupo de fazendas com 50, 50,1 e 50,2 acres não participarão do programa, por terem caído do lado errado do limiar de corte. O grupo de fazendas com 49,9 acres tem chances de ser muito semelhante ao grupo de fazendas com 50,1 acres em todos os aspectos, exceto que um grupo recebeu o subsídio de fertilizantes e o outro não. Conforme nos afastemos do ponto de corte, as unidades elegíveis e não elegíveis tornar-se-ão mais distintas por construção, porém temos uma medida de quão diferentes as fazendas são com base no critério de elegibilidade, o que nos permite controlar tais diferenças.

Uma vez que o programa é implementado e começa a subsidiar o custo de fertilizantes para pequenos e médios produtores, os avaliadores do programa podem usar um modelo de regressão descontínua para avaliar seu impacto. A regressão descontínua mede a diferença nos resultados pós-intervenção, tais como as produções totais de arroz, entre as unidades próximas ao ponto de corte de elegibilidade -50 acres, em nosso exemplo. As fazendas grandes demais para serem inscritas no programa constituem o grupo de comparação e geram uma estimativa do resultado contrafactual para as fazendas no grupo de tratamento que eram pequenas o suficiente para serem inscritas. Visto que os dois grupos de fazendas

eram muito semelhantes na linha de base e são expostos ao mesmo conjunto de fatores externos ao longo do tempo (como clima, choques de preços, políticas agrícolas locais e nacionais, etc.), a única explicação plausível para resultados diferentes no período pós-intervenção deve ser o programa.

O método de regressão descontínua nos permite estimar com sucesso o impacto de um programa, sem excluir nenhuma população elegível. No entanto, note que o impacto estimado somente é válido na vizinhança do ponto de corte de elegibilidade. No nosso exemplo, temos uma estimativa válida sobre o impacto do programa de subsídio para fertilizante para as maiores fazendas de médio porte - isto é, aquelas logo abaixo de 50 acres. A avaliação de impacto não necessariamente será capaz de identificar diretamente o impacto do programa em fazendas pequenas - aquelas, digamos, com 1 ou 2 acres de terra, onde os efeitos de um subsídio para fertilizantes podem diferir de modo relevante dos efeitos observados em fazendas de médio porte, com 48 ou 49 acres.

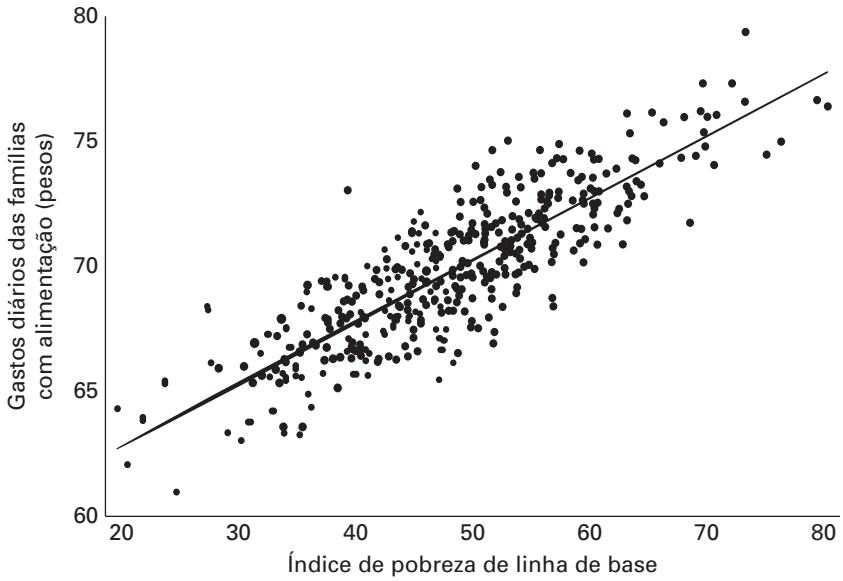
Não existe grupo de comparação para as fazendas pequenas, uma vez que todas elas são elegíveis para se inscrever no programa. A única comparação válida é para fazendas situadas perto do ponto de corte, de 50 acres.

Caso 2: Transferências de Renda

Vamos supor que estejamos tentando avaliar o impacto de um programa de transferência de renda nos gastos diários de alimentação das famílias pobres. Suponhamos, também, que possamos usar um índice de pobreza[1] que considera os ativos de uma família e os resume em um escore de 0 a 100 (onde valores mais próximos de 0 indicam maior nível de pobreza), usado para ranquear as famílias da mais pobre a mais rica. Na linha de base, você espera que as famílias mais pobres gastem menos em alimentos, em média, que as mais ricas. A figura 5.2 apresenta uma possível relação entre o índice de pobreza e os gastos diários das famílias (o resultado) com alimentos.

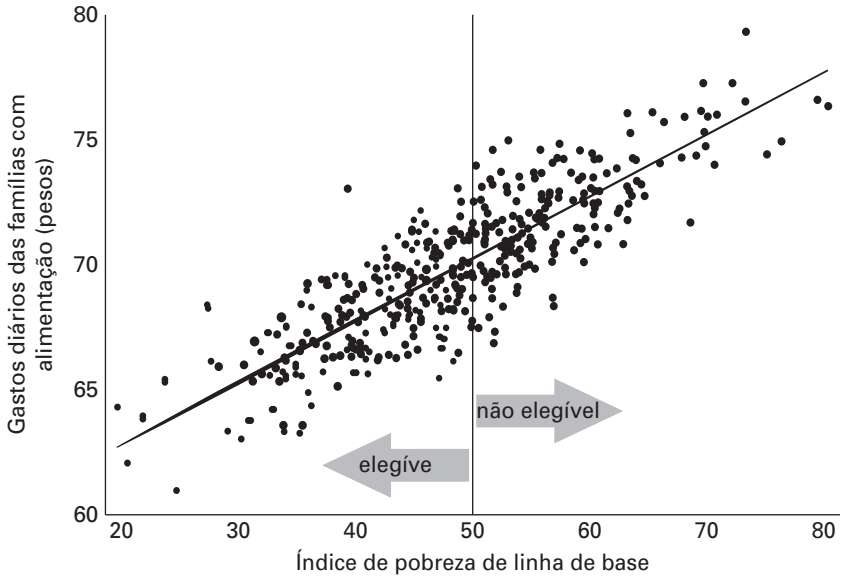
Suponhamos agora que o programa focalize somente nas famílias pobres, que são aquelas com um escore inferior a 50. Em outras palavras, o índice de pobreza pode ser usado para determinar a elegibilidade: será oferecido tratamento somente às famílias com um escore de pobreza igual ou inferior a 50. As famílias com um escore acima de 50 são inelegíveis. Neste exemplo, o índice de elegibilidade é simplesmente o índice de pobreza e o ponto de corte é o valor de 50. A relação contínua entre o índice de elegibilidade e a variável de resultado (gastos diários com alimentos) é ilustrada na figura 5.3. As famílias logo abaixo do limite de

Figura 5.2 Gastos das famílias em relação à pobreza (pré-intervenção)



Fonte: Autores.

Figura 5.3 Descontinuidade na elegibilidade para o programa de transferência de renda



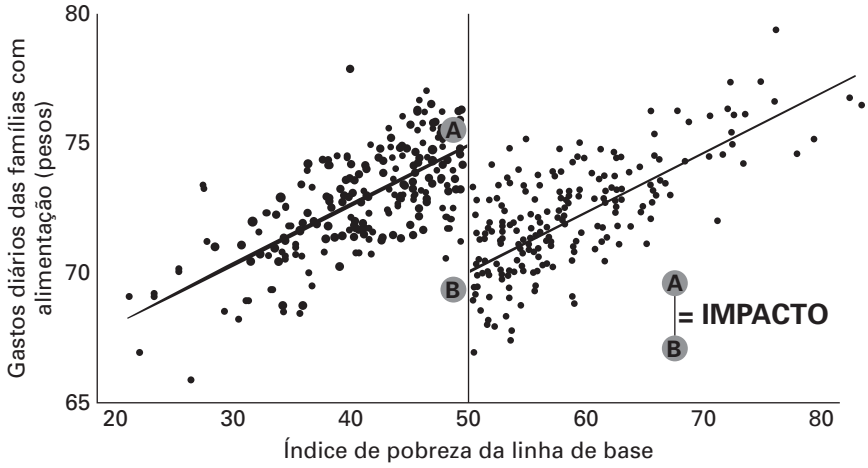
Fonte: Autores.

corte são elegíveis para o programa, enquanto que aquelas logo acima do limite de corte são inelegíveis, embora os dois tipos de famílias sejam muito semelhantes.

A estratégia RDD explora a descontinuidade em torno do ponto de corte para estimar o cenário contrafactual. Intuitivamente, as famílias um pouco abaixo do ponto de corte (logo abaixo de 50) são muito semelhantes àquelas com um escore um pouco acima do corte (por exemplo, 51). Usando o índice de pobreza, o programa decidiu um ponto particular (50) no qual há uma mudança repentina, ou descontinuidade, na elegibilidade para o programa. Uma vez que as famílias um pouco acima do ponto de corte de 50 são semelhantes àquelas um pouco abaixo (exceto por não receberem as transferências de renda), as famílias um pouco acima podem ser usadas como um grupo de comparação às famílias um pouco abaixo. Em outras palavras, as famílias que não são elegíveis, mas que estão muito próximas ao ponto de corte, serão usadas como grupo de comparação, para estimar o cenário contrafactual (o que teria acontecido ao grupo de famílias elegíveis na ausência do programa).

A figura 5.4 apresenta uma possível situação de pós-intervenção, transmitindo a intuição por trás da estratégia de identificação RDD. Os resultados médios das famílias (elegíveis) com níveis de pobreza de linha de base abaixo do escore de corte são, agora, mais altos doque os resultados médios das famílias (não elegíveis) com escores de linha de base logo

Figura 5.4 Gastos das Famílias em Relação à Pobreza (Pós-intervenção)



Fonte: Autores.

acima do corte. Dada a relação contínua entre o índice de pobreza e os gastos diários com alimentos antes do programa, a única explicação plausível para a descontinuidade observada após a intervenção deve ser a existência do programa de transferência de renda. Em outras palavras, como as famílias na vizinhança (direita e esquerda) do ponto de corte têm características semelhantes, a diferença na média de gastos com alimentos entre os dois grupos é uma estimativa válida do impacto do programa.

Usando o Método de Regressão Descontínua para Avaliar o Programa de Subsídio ao Seguro Saúde

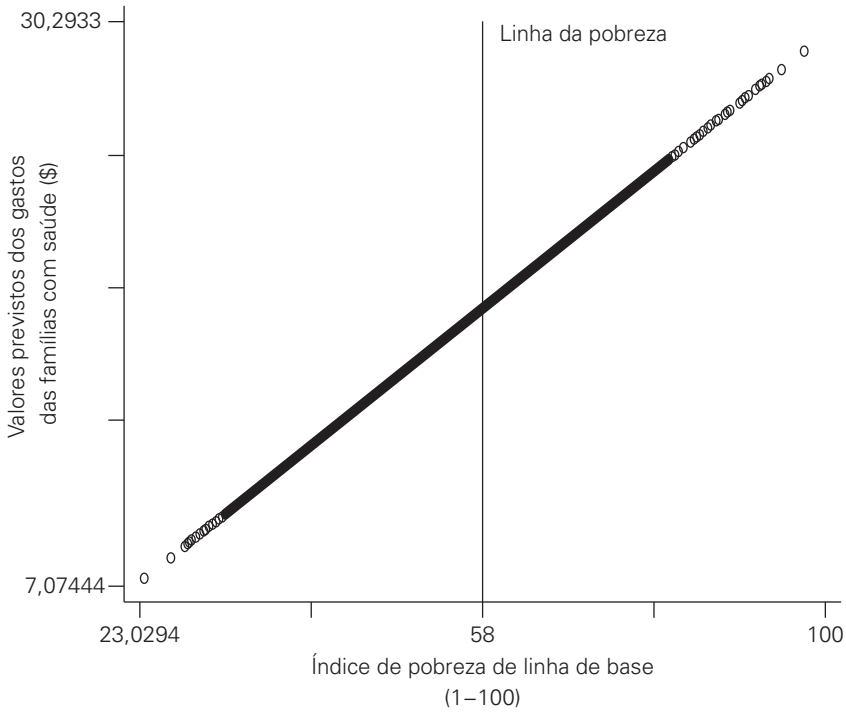
Apliquemos a regressão descontínua ao nosso Programa de Subsídio ao Seguro Saúde (HISP). Após algumas investigações sobre o desenho do HISP, você descobre que, na prática, as autoridades focalizaram o programa nas famílias de baixa renda usando a linha de pobreza nacional. A linha de pobreza baseia-se em um índice de pobreza, que atribui a cada família no país um escore entre 20 e 100, com base em seus ativos, condições de moradia e estrutura sociodemográfica.

A linha da pobreza foi estabelecida oficialmente na marca de 58. Isso significa que todas as famílias com uma pontuação inferior a 58 são classificadas como pobres e todas as famílias com uma pontuação superior a 58 não são consideradas pobres. Mesmo nos municípios de tratamento, somente as famílias pobres são elegíveis para se inscrever no HISP; todavia, a amostra inclui dados tanto das famílias pobres quanto daquelas em boa situação econômica nos municípios de tratamento.

Usando os domicílios dos municípios de tratamento de sua amostra, um colega lhe ajuda a realizar uma regressão multivariada e a representar graficamente a relação entre o índice de pobreza e os gastos previstos com saúde antes do início do HISP (figura 5.5). A figura demonstra claramente que, à medida que aumenta o escore das famílias no índice de pobreza, a regressão prevê um nível mais alto de gastos com saúde, refletindo o fato de que as famílias mais ricas tendem a ter maiores gastos e consumo de remédios e serviços básicos de saúde. Observe que a relação entre o índice de pobreza e os gastos com saúde é contínua - isto é, não há evidência de mudança na relação em torno da linha de pobreza.

Dois anos após o início do piloto, você observa que apenas as famílias com escore abaixo de 58 (isto é, à esquerda da linha de pobreza) puderam se

Figura 5.5 Índice de Pobreza e Gastos com Saúde na Linha de Base do Programa de Subsídio ao Seguro Saúde

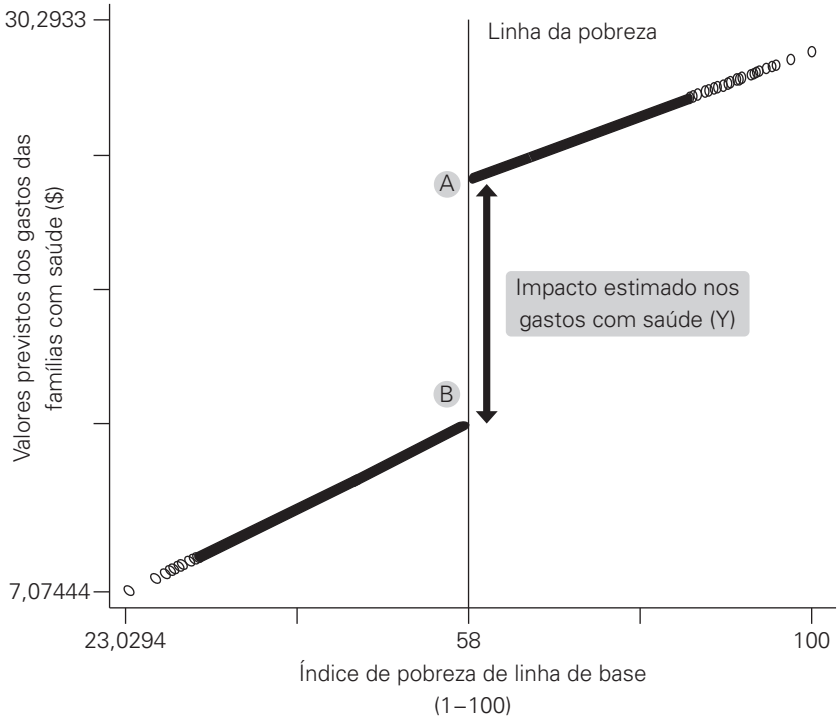


Fonte: Autores.

inscrever no HISP. Usando dados de seguimento, você novamente plota o gráfico da relação entre os escores no índice de pobreza e os gastos previstos com saúde, encontrando a relação apresentada na figura 5.6. Desta vez, a relação entre o índice de pobreza e os gastos previstos com a saúde não é mais contínua, existe uma clara interrupção, ou "descontinuidade", na linha de pobreza.

A descontinuidade reflete uma queda nos gastos com saúde entre as famílias elegíveis a receber o programa. Visto que as famílias em ambos os lados do limite de 58 pontos no escore de pobreza são muito semelhantes, a única explicação plausível para a diferença no nível de gastos com saúde é que um grupo de famílias era elegível para se inscrever no programa enquanto o outro, não. Estima-se essa diferença por meio de uma regressão com os achados apresentados na tabela 5.1.

Figura 5.6 Índice de Pobreza e Gastos com Saúde - Programa de Subsídio ao Seguro Saúde dois Anos Depois



Fonte: Autores.

Tabela 5.1 Caso 5 — Impacto do HISP Usando Regressão Descontínua (Análise de Regressão)

	Regressão Linear Multivariada
Impacto estimado nos	-9,05**
Gastos da família com saúde	(0,43)

Fonte: Autores.

Observação: Os erros padrão estão entre parênteses.

** Significativo a 1%.

PERGUNTA 5

A. O resultado apresentado na tabela 5.1 é válido para todas as famílias elegíveis?

B. Comparado ao impacto estimado usando alocação aleatória, o que este resultado diz sobre as famílias com um índice de pobreza pouco abaixo de 58?

C. Com base no resultado do caso 5, o HISP deveria ser estendido nacionalmente?

O Método RDD em Ação

O método da regressão descontínua tem sido utilizado em vários contextos. Lemieux e Milligan (2005) analisaram os efeitos da assistência social na disponibilidade de mão de obra no Quebec. Martinez (2004) estudou o efeito do valor das pensões dos idosos no consumo na Bolívia. Filmer e Schady (2009) avaliaram o impacto de um programa de concessão bolsas a estudantes pobres, para estimular a matrícula e aumentar as notas nos testes de aprendizagem no Camboja. Buddelmeyer e Skoufias (2004) examinaram o desempenho da regressão descontínua em relação ao experimento aleatório no caso do *Progresa* e descobriram que os impactos estimados usando os dois métodos são semelhantes para uma grande parte dos resultados analisados. Alguns destes exemplos são descritos detalhadamente nos quadros 5.1, 5.2 e 5.3.

Limitações e Interpretação do Método de Regressão Descontínua

A regressão descontínua estima impactos médios *locais* em torno do corte de elegibilidade no ponto onde as unidades de tratamento e de comparação

Quadro 5.1: Assistência Social e a Oferta de Trabalho no Canadá

Um dos estudos clássicos nos quais o método RDD foi utilizado se valeu de uma forte descontinuidade em um programa de assistência social no Quebec, no Canadá, para entender os efeitos do programa nos resultados do mercado de trabalho. O programa de assistência social, financiado pelo *Canadian Assistance Plan*, oferece ajuda aos desempregados. Durante muitos anos, o programa ofereceu pagamentos significativamente mais baixos a indivíduos com menos de 30 anos de idade e sem filhos, comparado aos indivíduos acima de 30 anos de idade - $185 por mês contra $507.

Para avaliar rigorosamente este programa, Lemieux e Milligan (2005) limitaram a amostra a homens sem filhos e sem diploma de nível superior e reuniram dados do Censo canadense e da Pesquisa sobre a Força de Trabalho (*Labor Force Survey*). Para justificar o uso da abordagem RDD, eles demonstraram que os homens próximos à descontinuidade (entre 25 e 39 anos de idade) eram muito semelhantes em características observáveis.

Comparando os homens de ambos os lados do limiar de elegibilidade, os autores descobriram que o acesso a benefícios maiores de assistência social reduziu, de fato, o emprego, em 4,5% para homens sem filhos naquela faixa etária.

Fonte: Lemieux e Milligan 2005.

Quadro 5.2: Mensalidades Escolares e Taxas de Matrícula na Colômbia

Na Colômbia, Barrera-Osorio, Linden e Urquiola (2007) usaram a regressão com descontinuidade para avaliar o impacto de um programa de redução da mensalidade escolar (*Gratuitad*) nas taxas de matrícula escolar na cidade de Bogotá. O programa é focalizado com base em um índice chamado SISBEN, que é um índice de pobreza contínuo, cujo valor é determinado por características domiciliares tais como a localização, materiais de construção da casa, serviços disponíveis, características demográficas, saúde, educação, renda, e ocupações dos membros da família. O governo estabeleceu dois pontos de corte ao longo do índice SISBEN: crianças de famílias com escores abaixo do ponto de corte n.º 1 são elegíveis para educação gratuita da 1ª a 11ª série; crianças de famílias com escores entre os pontos de corte n.º 1 e n.º 2 são elegíveis para um subsídio de 50% nas mensalidades da 10ª e 11ª séries; e crianças de famílias com escores acima do ponto de corte n.º 2 não são elegíveis para educação gratuita ou subsídios.

Os autores usaram a regressão com descontinuidade por quatro razões. Primeiro, as características familiares do chefe da família, como renda ou nível educacional, são contínuas ao longo do escore SISBEN na linha de base; em outras palavras, não há "saltos" nas características ao longo do escore SISBEN. Em segundo lugar, as famílias em ambos os lados dos pontos de corte possuem características semelhantes, sugerindo que o método produziu grupos de comparação confiáveis. Em terceiro lugar, uma grande amostra de famílias estava disponível. Finalmente, o governo manteve a fórmula usada para calcular o índice SISBEN em segredo, a fim de que as famílias não pudessem manipular suas pontuações.

Usando o método RDD, os pesquisadores descobriram que o programa tinha um significante impacto positivo nas taxas de matrícula escolar. Especificamente, a matrícula era três pontos percentuais maior para alunos do Ensino Fundamental de famílias abaixo do ponto de corte n.º 1 e 6% maior para alunos do Ensino Médio das famílias entre o ponto de corte n.º 1 e n.º 2. O estudo fornece evidência sobre os benefícios de reduzir os custos diretos de escolaridade, particularmente para alunos em situação de risco. No entanto, os autores apontam a necessidade de mais pesquisas sobre a elasticidade de preços, para que se informe melhor o desenho de programas de subsídio tais como este.

Fonte: Barrera-Osorio, Linden, e Urquiola 2007.

mais se assemelham. Conforme nos aproximamos do corte, as unidades que ficam à esquerda e à direita dele se parecem mais. De fato, quando chegamos bem perto do ponto de corte, as unidades à esquerda e à direita da linha são tão semelhantes que a nossa comparação será tão boa quanto se tivéssemos escolhido os grupos de tratamento e de comparação usando a alocação aleatória de tratamento.

Quadro 5.3: Redes de Proteção Social com base em um Índice de Pobreza na Jamaica

O método RDD também foi usado para avaliar o impacto de uma iniciativa de rede de proteção social na Jamaica. Em 2001, o governo jamaicano iniciou o Programa de Avanço Através da Saúde e Educação (PATH) para aumentar os investimentos no capital humano e melhorar a focalização dos benefícios de bem-estar social fornecidos aos pobres. O programa oferecia transferências monetárias para crianças de famílias pobres elegíveis, condicionadas à frequência escolar e visitas regulares ao serviço de saúde. O benefício médio mensal para cada criança era de $6,50, além da isenção de certas taxas de saúde e educação.

Dado que a elegibilidade para o programa foi determinada por uma fórmula de pontuação, Levy e Ohls (2007) puderam comparar as famílias que estavam um pouco abaixo do limiar de elegibilidade àquelas um pouco acima (entre 2 e 15 pontos do corte). Os pesquisadores justificaram o uso do método RDD com dados de linha de base mostrando que os domicílios de tratamento e comparação tinham níveis semelhantes de pobreza, medida por uma variável indireta (*proxy)* das condições de vida e níveis semelhantes de motivação, já que todas as famílias na amostra se inscreveram no programa. Os pesquisadores também usaram o escore de elegibilidade do programa na análise de regressão para ajudar a controlar diferenças entre os grupos.

Levy e Ohls (2007) descobriram que o programa PATH aumentou o índice de frequência escolar de crianças com idades entre 6 e 17 anos em uma média de meio dia por mês – algo significativo em face a uma já elevada taxa de frequência de 85%. De forma semelhante, as visitas médicas das crianças de 0 a 6 anos de idade aumentaram em aproximadamente 38%. Apesar dos pesquisadores não encontrarem impactos de longo prazo no desempenho escolar dos alunos ou em seu estado de saúde, eles concluíram que a magnitude dos resultados encontrados era em geral consistente com os demais programas de transferência condicionada de renda implementados em outros países. Um último aspecto interessante desta avaliação foi reunir dados tanto quantitativos quanto qualitativos, usando sistemas de informação, entrevistas, grupos focais e pesquisas domiciliares.

Fonte: Levy e Ohls 2007.

Dado que o método RDD estima o impacto do programa em torno do ponto de corte (ou seja, *localmente*), a estimativa não necessariamente pode ser generalizada às unidades cujos escores estão mais distantes do ponto de corte - isto é, onde os indivíduos elegíveis e não elegíveis podem não ser tão semelhantes. O fato de que o método RDD não poderá computar um efeito médio do tratamento para todos os participantes do programa pode ser visto tanto como um ponto forte quanto uma limitação do método,

dependendo da pergunta de interesse da avaliação. Se a avaliação busca primeiramente responder à pergunta "O programa deveria ou não existir?", então o efeito médio do tratamento sobre toda a população elegível pode ser o parâmetro mais relevante e claramente a RDD fica aquém de ser perfeita. No entanto, se a pergunta de interesse for "O programa deveria ser interrompido ou expandido na margem?", então a RDD produz precisamente a estimativa local de interesse para informar esta importante decisão de política.

O fato de que o método RDD estima o efeito médio de tratamento local também aumenta os desafios em termos do poder estatístico da análise. Uma vez que os efeitos são estimados somente em torno do ponto de corte, menos observações podem ser usadas em comparação a outros métodos, que incluiriam todas as unidades de observação. São necessárias amostras de avaliação relativamente grandes para que se obtenha poder estatístico suficiente ao aplicar a RDD. Na prática, determinamos a largura da banda em torno do ponto de corte que será incluída na estimação, considerando o equilíbrio nas características observadas da população acima e abaixo do corte. Poderemos, então, refazer a estimativa usando diferentes larguras de banda para verificar se as estimativas são sensíveis à largura de banda escolhida. Como regra geral, quanto maior a largura da banda, maior será o poder estatístico da análise, já que mais observações são incluídas na amostra. No entanto, afastar-se do ponto de corte pode, também, exigir hipóteses adicionais a respeito da forma funcional para que se obtenha uma estimativa confiável do impacto.

Uma advertência adicional ao usar o método RDD é que a especificação pode ser sensível à forma funcional usada na modelação da relação entre o escore de elegibilidade e o resultado de interesse. No exemplo do programa de transferência de renda, assumimos que a relação de linha de base entre o índice de pobreza das famílias e os seus gastos diários em alimentos fosse simplesmente linear. Na realidade, a relação entre o índice de elegibilidade e o resultado de interesse (Y) na linha de base poderia ser muito mais complexa e poderia envolver relações não lineares e interações entre as variáveis. Se não levarmos em conta essas relações complexas na estimativa, elas podem, por engano, ser tomadas como uma descontinuidade nos resultados pós-intervenção. Na prática, podemos estimar o impacto do programa usando várias formas funcionais (linear, quadrada, cúbica, etc.) para avaliar se, de fato, as estimativas de impacto são sensíveis à forma funcional.

Mesmo com estas limitações, a regressão descontínua gera estimativas não enviesadas do impacto na vizinhança do ponto de corte de elegibilidade.

A estratégia de regressão descontínua se beneficia das regras de alocação do programa, usando índices de elegibilidade contínua comuns em muitos

programas sociais. Quando são aplicadas regras de focalização com base em índices, não é necessário excluir um grupo de famílias ou indivíduos elegíveis do tratamento para fins de avaliação, uma vez que a regressão com descontinuidade pode ser usada.

Nota

1. Isto é chamado, às vezes, de "teste de elegibilidade multidimensional", porque leva em conta os ativos das famílias como um indicador ou estimador da condição de vida ou do poder de compra.

Referências

Barrera-Osorio, F., Linden, L. & Urquiola, M. (2007). The Effects of User Fee Reductions on Enrollment: Evidence from a Quasi-Experiment. Washington, DC: Columbia University e Banco Mundial.

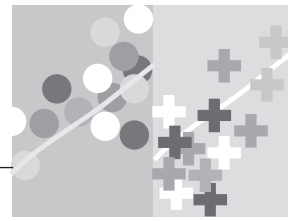
Buddelmeyer, H., & Skoufias, E. (2004). An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA. *Documento de Trabalho de Pesquisa em Política do Banco Mundial 3386, IZA Documentos de Discussão 827*. Washington, DC: Banco Mundial.

Filmer, D. & Schady, N. (2009). School Enrollment, Selection and Test Scores. *Documento de Trabaho de Pesquisa em Política do Banco Mundial 4998*. Washington, DC: Banco Mundial.

Lemieux, T. & Milligan, K. (2005). Incentive Effects of Social Assistance: a Regression Discontinuity Approach. *Documento de Trabalho NBER 10541*. Cambridge, MA: Escritório Nacional de Pesquisa Econômica

Levy, D. & Ohls, J. (2007). Evaluation of Jamaica's PATH Program: Final Report. *Mathematica Policy Research, Inc., Ref. 8966-090*, Washington, DC.

Martinez, S. (2004). Pensions, Poverty and Household Investments in Bolivia. Berkeley, CA: Universidade da Califórnia.



CAPÍTULO 6

Diferença-em-Diferenças

Os três métodos de avaliação de impacto discutidos até agora - *seleção ou alocação aleatória, promoção aleatória,* e *regressão descontínua* (RDD) - todos produzem estimativas do contrafactual através de regras explícitas de alocação do programa, que o avaliador conhece e entende. Discutimos por que esses métodos oferecem estimativas confiáveis do contrafactual com relativamente poucos suposições e condicionantes. Os próximos dois tipos de métodos – *diferença-em-diferenças* (DD) e *métodos de pareamento* - oferecem ao avaliador um conjunto adicional de ferramentas que podem ser aplicadas em situações nas quais as regras de seleção do programa estejam menos claras, ou nas quais nenhum dos três métodos descritos previamente seja factível. Conforme veremos, tanto o método DD quanto o de pareamento podem ser poderosas ferramentas estatísticas; muitas vezes eles serão usados juntos ou em combinação com outros métodos de avaliação de impacto.

Tanto o método de *diferença-em-diferenças* quanto o de pareamento são comumente usados; no entanto, ambos exigem mais suposições do que os métodos de seleção aleatória. Também chamamos atenção desde já ao fato de que ambos os métodos demandam a existência de dados de linha de base.[1]

O método de *diferença-em-diferenças* faz o que seu nome sugere. Compara as *mudanças* nos resultados ao longo do tempo em uma população inscrita em um programa (o grupo de tratamento) e uma população que não está

Conceito-chave:
O método *diferença-em-diferenças* estima o contrafactual para a mudança no resultado do grupo de tratamento ao calcular a variação no resultado do grupo de comparação. Este método permite levar em conta diferenças entre os grupos de tratamento e de comparação que sejam constantes ao longo do tempo.

inscrita (o grupo de comparação). Consideremos, por exemplo, um programa de construção de estradas que não pode ser alocado aleatoriamente e não é alocado com base em um índice com um ponto de corte claramente definido, que permitiria a utilização de uma regressão descontínua (RDD). Um dos objetivos do programa é melhorar o acesso ao mercado de trabalho, sendo um dos indicadores de resultado o nível de emprego.

Conforme vimos no capítulo 3, a simples observação de mudança entre o antes e o depois nas taxas de emprego para as áreas afetadas pelo programa não nos dará o impacto causal do programa, visto que muitos outros fatores podem influenciar o nível de emprego ao longo do tempo. Ao mesmo tempo, comparar áreas que receberam o programa de estradas às que não o receberam será problemático se existirem razões não observadas para que algumas áreas recebessem o programa enquanto outras não (o problema do viés de seleção discutido no cenário dos inscritos versus os não inscritos).

No entanto, e se combinarmos os dois métodos e compararmos as mudanças ocorridas entre o antes e o depois nos resultados de um grupo que se inscreveu no programa com as mudanças entre o antes e o depois de um grupo que não se inscreveu no programa? A diferença nos resultados entre o antes e o depois para o grupo inscrito - a primeira diferença – *controla* os fatores constantes ao longo do tempo naquele grupo, uma vez que estamos comparando o grupo com ele mesmo. Mas ainda nos faltam os fatores externos que variam com o tempo. Uma forma de captar os fatores que variam ao longo do tempo é medir a mudança entre o antes e o depois nos resultados para um grupo que *não* se inscreveu no programa, mas foi exposto ao mesmo conjunto de condições ambientais - a segunda diferença. Se "limparmos" a primeira diferença dos outros fatores que variam com o tempo e afetam o resultado de interesse subtraindo a segunda diferença, então teremos eliminado a principal fonte do viés que nos preocupava nas comparações do tipo *antes e depois*. A abordagem *diferença-em-diferenças* combina, portanto, os dois cenários contrafactuais falsos (comparações *antes-e-depois* e comparações entre os que escolherem se inscrever e os que decidiram não se inscrever) para produzir uma melhor estimativa do cenário contrafactual. No exemplo das estradas, o método DD poderia comparar a mudança no nível de emprego antes e depois da implementação do programa para indivíduos que moram nas áreas afetadas pelo programa de construção de estradas com a mudança no emprego em áreas onde o programa de estradas não foi implementado.

É importante observar que o cenário contrafactual que está sendo estimado aqui é a *mudança* nos resultados para o grupo de comparação. Os grupos de tratamento e de comparação não precisam, necessariamente, ter as mesmas condições de pré-intervenção. Porém, para que o método DD seja

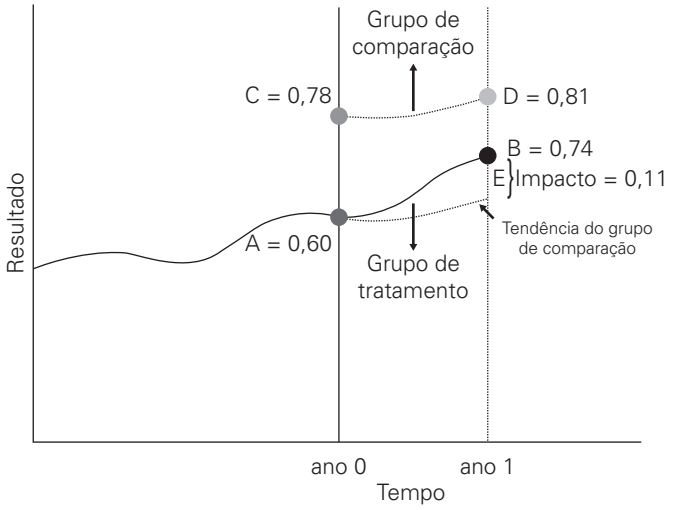
válido, o grupo de comparação deve representar precisamente a mudança nos resultados que seriam experimentados pelo grupo de tratamento na falta do tratamento. Para aplicar o *diferença-em-diferenças*, tudo o que é preciso é medir os resultados no grupo que recebe o programa (o grupo de tratamento) e o grupo que não o recebe (o grupo de comparação), tanto antes quanto depois do programa. O método não requer que especifiquemos as regras pelas quais o tratamento é atribuído.

A figura 6.1 ilustra o método *diferença-em-diferenças*. O grupo de tratamento está inscrito em um programa e o grupo de comparação não está inscrito.

As variáveis de resultado antes e depois para o grupo de tratamento são *A* e *B,* respectivamente, enquanto o resultado para o grupo de comparação vai do *C,* antes do programa, ao *D*, após a implementação do programa.

Você se lembrará dos nossos dois contrafactuais falsos - a diferença nos resultados antes e depois da intervenção para o grupo de tratamento *(B - A)* e a diferença nos resultados[2] após a intervenção, entre os grupos de tratamento e de comparação *(B - D)*. No modelo *diferença-em-diferenças,* a estimativa do cenário contrafactual é obtida ao calcular-se a mudança nos resultados no grupo de comparação *(D - C)*. Esta mudança contrafactual é, então, subtraída da mudança nos resultados no grupo de tratamento *(B - A)*.

Figura 6.1 ***Diferença-em-Diferenças***



Fonte: Autores.

Em suma, o impacto do programa é calculado, simplesmente, pela diferença entre duas diferenças:

Impacto DD = *(B − A)* − *(D − C)* = *(B − E)* = (0,74 – 0,60) − (0,81 – 0,78) = 0,11.

As relações apresentadas na figura 6.1 também podem ser exibidas em uma tabela simples. A Tabela 6.1 separa os componentes das estimativas de *diferença-em-diferenças*. A primeira linha contém resultados para o grupo de tratamento antes *(A)* e depois *(B)* da intervenção. A comparação antes e depois para o grupo de tratamento é a primeira diferença *(B - A)*. A segunda linha contém resultados do grupo de comparação antes da intervenção *(C)* e depois da intervenção *(D)*; assim, a segunda diferença (contrafactual) é *(D - C)*.

O método *diferença-em-diferenças* calcula a estimativa de impacto da seguinte forma:

1. Calculamos a diferença no resultado *(Y) entre* as situações antes e depois para o grupo de tratamento *(B - A)*.
2. Calculamos a diferença no resultado *(Y)* entre as situações antes e depois para o grupo de comparação *(D - C)*.
3. Então calculamos a diferença entre a diferença nos resultado para o grupo de tratamento *(B - A)* e a diferença para o grupo de comparação *(D - C)*, ou DD = *(B - A)* - *(D - C)*. Esta "diferença-em-diferenças" é a nossa estimativa de impacto.

Tabela 6.1 O Método *Diferença-em-Diferenças*

	Depois	**Antes**	**Diferença**
Tratamento/inscritos	*B*	*A*	*B – A*
Comparação/ não inscritos	*D*	*C*	*D – C*
Diferença	*B – D*	*A – C*	*DD = (B – A) – (D – C)*

	Depois	**Antes**	**Diferença**
Tratamento/inscritos	0,74	0,60	*0,14*
Comparação/ não inscritos	0,81	0,78	*0,03*
Diferença	*–0,07*	*–0,18*	*DD = 0,14 – 0,03 = 0,11*

Fonte: Autores.

Como o Método *Diferença-em-Diferenças* pode ser Útil?

Para entender como o método *diferença-em-diferenças* pode ajudar, vamos começar pelo nosso segundo contrafactual falso, que comparou unidades que se inscreveram no programa àquelas que não se inscreveram. Lembre-se que a principal preocupação foi o fato dos dois conjuntos de unidades poderem ter características diferentes, que explicariam (ao invés do programa) a diferença nos resultados entre os dois grupos. As diferenças em características *não observadas* eram particularmente preocupantes: por definição, é impossível incluir diferenças em características não observadas na análise.

O método *diferença-em-diferenças* ajuda a resolver esse problema, na medida em que muitas características das unidades ou indivíduos podem ser razoavelmente assumidas como constantes ao longo do tempo (ou *invariantes no tempo*). Pense, por exemplo, em características *observáveis*, como o ano de nascimento de uma pessoa, a localização próxima ao oceano de uma região, o nível de desenvolvimento econômico de uma cidade, ou o nível de educação de um pai. A maioria dessas variáveis, embora possivelmente relacionadas aos resultados, provavelmente não mudarão no decorrer de uma avaliação. Usando o mesmo raciocínio, podemos concluir que muitas características *não observadas* dos indivíduos são mais ou menos constantes ao longo do tempo. Considere, por exemplo, a inteligência de uma pessoa ou seus traços de personalidade, como motivação, otimismo, autodisciplina, ou histórico familiar de saúde. É possível que muitas dessas características intrínsecas a uma pessoa não mudem ao longo do tempo.

Quando o mesmo indivíduo é observado antes e depois de um programa e computamos uma diferença simples no resultado para este indivíduo, cancelamos o efeito de todas as características que são exclusivas a esse indivíduo e que não mudam ao longo do tempo. O interessante é que estaremos cancelando (ou controlando) não somente o efeito das características *observadas* que não variam com o tempo, mas também o efeito das características *não observadas* que não variam com o tempo, tais como as mencionadas acima.

O Suposição da "Igualdade de Tendências" no Método Diferença-em-Diferenças

Embora o método *diferença-em-diferenças* nos permita tratar das diferenças entre os grupos de tratamento e de comparação que são constantes ao longo do tempo, ele não nos ajudará a eliminar as diferenças entre os grupos de tratamento e de comparação que mudam ao longo do tempo. No exemplo

das estradas acima, se as áreas de tratamento também se beneficiam da construção de um novo porto marítimo ao mesmo tempo em que se beneficiam da construção da estrada, não conseguiremos levar em conta a construção do porto nas nossas estimativas ao usar a abordagem *diferença-em-diferenças*. Para que o método forneça uma estimativa válida do contrafactual, precisamos assumir que não existem diferenças que variem com o tempo entre os grupos de tratamento e de comparação.

Outra maneira de se pensar nisso é que, na ausência do programa, as diferenças nos resultados entre os grupos de tratamento e de comparação precisariam se mover conjuntamente. Isto é, sem o tratamento, os resultados precisariam aumentar ou diminuir na mesma taxa nos dois grupos; é necessário que os resultados apresentem *tendências iguais na ausência do tratamento*.

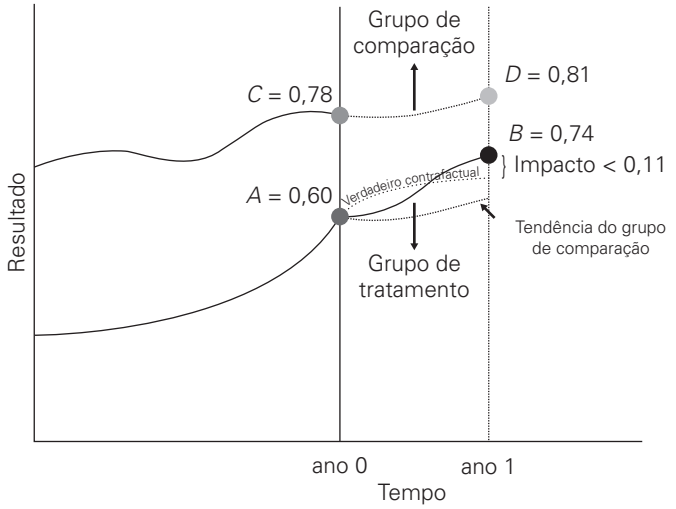
Infelizmente, não há forma de provarmos que as diferenças nos resultados entre os grupos de tratamento e de comparação teriam se movido paralelamente na ausência do programa. A razão é que não podemos observar o que teria acontecido ao grupo de tratamento na ausência do programa - em outras palavras, não podemos observar o contrafactual!

Portanto, quando usamos o método *diferença-em-diferenças*, devemos *supor* que, na ausência do programa, o resultado do grupo de tratamento teria se deslocado em paralelo ao resultado do grupo de comparação. A figura 6.2 ilustra a violação deste pressuposto fundamental, necessário para que o método *diferença-em-diferenças* produza estimativas confiáveis de impacto. Se as tendências de resultados para os grupos de tratamento e de comparação são diferentes, então o efeito do tratamento estimado obtido pelo método *diferença-em-diferenças* seria inválido ou enviesado. O motivo é que a tendência observada para o grupo de comparação não é uma estimativa válida da tendência contrafactual que teria prevalecido para o grupo de tratamento na ausência do programa. Conforme vemos na figura 6.2, os resultados para o grupo de comparação crescem mais rapidamente que os resultados para o grupo de tratamento, na ausência do programa. Sendo assim, usar a tendência do grupo de comparação como contrafactual da tendência do grupo de tratamento leva a uma sobrestimação do impacto do programa.

Testando a Validade da Hipótese de "Igualdade de Tendências" no Modelo de Diferença-em-Diferenças

A validade da hipótese subjacente da igualdade de tendências pode ser avaliada, embora não possa ser provada. Um bom teste da validade dessa hipótese é comparar as mudanças nos resultados para os grupos de tratamento e de comparação *antes* da implementação do programa. Se as variáveis de

Figura 6.2 ***Diferença-em-Diferenças*** **quando as Tendências dos Resultados diferem**



Fonte: Autores.

resultados se moviam juntos (para o grupo de tratamento e de comparação) antes do início do programa, teremos maior confiança de que seguiriam a mesma tendência no período pós-intervenção. Para verificar a igualdade das tendências pré-intervenção, precisamos, pelo menos, de duas observações em série nos grupos de tratamento e comparação antes do início do programa. Isto quer dizer que a avaliação precisaria de três observações em série - duas observações pré-intervenção, para avaliar as tendências antes do programa e, pelo menos, uma observação pós-intervenção, para avaliar o impacto com a fórmula de *diferença-em-diferenças*.

Uma segunda maneira de testar a hipótese de igualdade de tendências seria executar o que é conhecido como o teste de "placebo". Neste teste, você executa uma estimativa adicional de *diferença-em-diferenças* usando um grupo de tratamento "falso", isto é, um grupo que você sabe que não foi afetado pelo programa. Digamos, por exemplo, que você estime como as aulas extras de reforço para alunos da sétima série afetam sua probabilidade de frequentar a escola e que você escolha os alunos da oitava série como o grupo de comparação. Para testar se os alunos da sétima e da oitava série possuem a mesma tendência em termos de frequência escolar, você poderia testar se os alunos da oitava e sexta séries possuem as mesmas tendências. Você sabe que o programa não afeta os alunos da sexta série, então se você executar uma estimação de *diferença-em-diferenças* usando os alunos da

oitava série como grupo de comparação e os alunos da sexta série como grupo de tratamento falso, você *terá* que encontrar um impacto zero. Caso não seja zero, então o impacto estimado deve vir de alguma diferença subjacente às tendências entre os alunos da sexta e da oitava série. Isto, por sua vez, põe em dúvida se os alunos da sétima e oitava séries podem ser considerados como tendo tendências paralelas na ausência do programa.

Um teste placebo pode ser realizado não somente com um grupo de tratamento falso, mas também com um resultado falso. No exemplo das aulas de reforço, você pode querer testar a validade de usar os alunos da oitava série como grupo de comparação, estimando o impacto das aulas de reforço em um resultado que você sabe que não é afetado por elas, tal como o número de irmãos que os alunos têm. Se a sua estimativa usando *diferença-em-diferenças* resultar em um "impacto" das aulas de reforço no número de irmãos que os alunos têm, então você saberá que o seu grupo de comparação deve ser falho.

Uma quarta maneira de se testar a hipótese das tendências paralelas seria aplicar o método de *diferença-em-diferenças* usando diferentes grupos de comparação. No exemplo das aulas de reforço, você faria primeiramente a estimativa usando os alunos da oitava série como grupo de comparação e, em seguida, faria uma segunda estimativa usando os alunos da sexta série como o grupo de comparação. Se ambos os grupos forem grupos válidos de comparação, você constatará que o impacto estimado é aproximadamente o mesmo em ambos os cálculos.

Usando o Método *Diferença-em-Diferenças* para Avaliar o Programa de Subsídio ao Seguro Saúde

O método *diferença-em-diferenças* pode ser usado para avaliar nosso Programa de Subsídio ao Seguro Saúde (HISP). Neste caso, você possui duas rodadas de dados de dois grupos de famílias: um grupo que se inscreveu no programa e outro que não se inscreveu. Lembrando o caso dos grupos selecionados de inscritos e não inscritos, você percebe que não pode simplesmente comparar os gastos médios com saúde dos dois grupos, devido ao viés de seleção. Uma vez que você possui dados de dois períodos para cada família da amostra, você pode usar esses dados para resolver alguns dos desafios comparando a mudança nos gastos para os dois grupos, assumindo que a mudança nos gastos com saúde do grupo de não inscritos reflita o que teria acontecido com os gastos do grupo inscrito, na ausência do programa (vide Tabela 6.2). Observe que não importa a direção em que se calcula a dupla diferença.

Tabela 6.2 Caso 6 — O Impacto do HISP Usando *Diferença-em-Diferenças* (Comparação de Médias)

	Depois (seguimento)	Antes (linha de base)	Diferença
Inscritos	7,8	14,4	–6,6
Não inscritos	21,8	20,6	1,2
Diferença			DD = –6,6 – 1,2 = –7,8

Fonte: Autores.

Tabela 6.3 Caso 6 — Impacto do HISP usando *diferença-em-diferenças* (análise de regressão)

	Regressão linear	Regressão linear multivariada
Impacto estimado nas	–7,8**	–7,8**
despesas da família com saúde	(0,33)	(0,33)

Fonte: Autores.

Observação: Os erros padrão se encontram entre parênteses.

** Significativo a 1%.

Em seguida, você estima o efeito usando uma análise de regressão (Tabela 6.3). Usando uma regressão linear simples, você descobre que o programa reduziu os gastos das famílias com saúde em $ 7,80. Você, então, refina a sua análise, usando uma regressão linear multivariada para considerar vários outros fatores e encontra a mesma redução nos gastos da família com saúde.

PERGUNTA 6 :

A. Quais são os pressupostos básicos necessários para se aceitar este resultado do caso 6?

B. Com base no resultado do caso 6, o HISP deveria ser estendido nacionalmente?

O Método *Diferença-em-Diferenças* em Ação

Apesar de suas limitações, o método de *diferença-em-diferenças* continua sendo um dos mais frequentemente utilizados dentre as metodologias de avaliação de impacto, com muitos exemplos encontrados na literatura.

Por exemplo, Duflo (2001) analisou os impactos da construção de escolas na Indonésia no nível de escolaridade e no mercado de trabalho. DiTella e Schargrodsky (2005) examinaram se o aumento das forças policiais reduziria o crime. Outro exemplo-chave da literatura está descrito no quadro 6.1.

Quadro 6.1: Privatização da Água e a Mortalidade Infantil na Argentina

Galiani, Gertler e Schargrodsky (2005) usaram o método de *diferença-em-diferenças* para abordar uma importante questão de política: se a privatização dos serviços de água poderia melhorar os resultados na saúde e ajudar a diminuir a pobreza. Durante a década de 90, a Argentina deu início a uma das maiores campanhas de privatização já realizadas, transferindo as empresas locais de água para empresas privadas reguladas, cobrindo cerca de 30% dos municípios do país e 60% da população. O processo de privatização levou mais de dez anos, com a maioria das privatizações ocorrendo após 1995.

Galiani, Gertler e Schargrodsky (2005) aproveitaram a variação no estado de propriedade ao longo do tempo para determinar o impacto da privatização na mortalidade de crianças com menos de 5 anos de idade. Antes de 1995, as taxas de mortalidade infantil vinham caindo quase no mesmo ritmo em toda a Argentina; após 1995, as taxas de mortalidade foram reduzidas com mais rapidez nos municípios que privatizaram os serviços de água. Os pesquisadores argumentam que, neste contexto, as hipóteses de identificação por trás da *diferença-em-diferenças* são, provavelmente, verdadeiras. Primeiro, eles mostram que a decisão de privatizar não estava correlacionada com choques econômicos ou os níveis históricos de mortalidade infantil. Em segundo lugar, eles mostraram que nenhuma diferença nas tendências de mortalidade infantil podia ser observada entre os municípios de comparação e de tratamento antes do começo do movimento de privatização.

Eles verificaram a validade de seus resultados ao decompor o efeito da privatização na mortalidade infantil pelas diferentes causas de morte e descobriram que a privatização dos serviços de água está correlacionada com as reduções nas mortes causadas por doenças infecciosas ou parasitárias, mas não por causas que não apresentavam relação às condições da água, tais como acidentes ou doenças congênitas. Por fim, a avaliação verificou que os índices de mortalidade infantil caíram 8% nas áreas que privatizaram os serviços de água e que o efeito foi aproximadamente 26% maior nas regiões mais pobres, onde a expansão da rede de água foi maior. Este estudo lançou luz sobre vários pontos do debate de políticas em torno da privatização dos serviços públicos. Os pesquisadores concluíram que, na Argentina, o setor privado regulado provou ser mais bem sucedido do que o setor público em melhorar os indicadores de acesso, serviço e, principalmente, mortalidade infantil.

Fonte: Galiani, Gertler, e Schargrodsky 2005.

Limitações do Método *Diferença-em-Diferenças*

O método *diferença-em-diferenças* é, geralmente, menos robusto que os métodos de seleção aleatória (alocação aleatória, oferta aleatória e promoção aleatória). Mesmo quando as tendências são paralelas antes da intervenção, o viés na estimativa ainda pode aparecer. A razão é que o DD atribui à intervenção *qualquer diferença nas tendências* entre os grupos de tratamento e de comparação que ocorram *a partir do momento em que a intervenção começa*. Se quaisquer outros fatores que afetem a diferença nas tendências entre os dois grupos estiverem presentes, a estimativa será inválida ou enviesada.

Digamos que você esteja tentando estimar o impacto na produção de arroz do subsídio aos fertilizantes e está fazendo isso medindo a produção de arroz dos agricultores subsidiados (tratamento) e não subsidiados (comparação) antes e depois da distribuição dos subsídios. Se no ano 1 os agricultores subsidiados foram afetados por secas, enquanto os agricultores não subsidiados não foram afetados, então, a estimativa *diferença-em-diferenças* produzirá uma estimativa inválida do impacto do subsídio aos fertilizantes. Em geral, qualquer fator que afete somente o grupo de tratamento e o faça ao mesmo tempo em que o grupo recebe o tratamento tem o potencial para invalidar ou enviesar a estimativa do impacto do programa. O método *diferença-em-diferenças assume* que tal fator não está presente.

Notas

1. Embora a alocação aleatória, a promoção aleatória e a regressão descontínua, teoricamente, não requeiram dados de linha de base, na prática, ter uma linha de base é muito útil para confirmar se as características dos grupos de tratamento e de comparação estão equilibradas. Por essa razão, recomendamos incluir uma linha de base como parte da avaliação. Além de verificar o equilíbrio das características, há várias outra razões para a coleta de dados de linha de base, mesmo quando o método absolutamente não a requer. Primeiramente, possuir características de uma população pré-intervenção (exógena) permite ao avaliador determinar se o programa tem um impacto diferente nos diferentes grupos da população elegível (conhecida por "análise de heterogeneidade"). Em segundo lugar, os dados de linha de base também são usados para executar análises que podem orientar a política mesmo antes do início da intervenção e coletar os dados de linha de base pode servir como um piloto massivo para a coleta de dados pós-intervenção. Em terceiro lugar, os dados de linha de base podem servir como uma "apólice de seguro", caso a alocação aleatória não seja implementada; como uma segunda opção, o avaliador poderia usar uma combinação de *pareamento* e *diferença-em-diferenças*. Finalmente, os dados de linha de base

podem adicionar poder estatístico à análise, quando o número de unidades nos grupos de tratamento e de comparação for limitado.

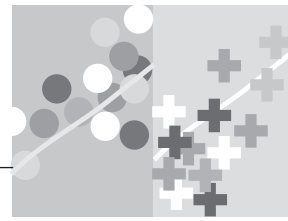
2. Todas as diferenças entre os pontos deveriam ser lidas como diferenças verticais dos resultados no eixo vertical.

Referências

DiTella, R., & Schargrodsky, E. (2005). Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack. *American Economic Review* 94 (1): 115–33.

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Review* 91 (4): 795–813.

Galiani, S., Gertler, P. & Schargrodsky, E. (2005). Water for Life: The Impact of the Privatization of Water Services on Child Mortality. *Journal of Political Economy* 113 (1): 83–120.



CAPÍTULO 7

Pareamento

O método descrito neste capítulo consiste em um conjunto de técnicas estatísticas às quais nos referiremos coletivamente como "pareamento". Os métodos de pareamento podem ser aplicados no contexto de quase todas as regras de alocação do programa, enquanto existir um grupo que não tenha participado do programa. Os métodos de pareamento geralmente dependem de características observadas para a construção de um grupo de comparação. Portanto, exigem suposição de que não existem diferenças não observadas nas populações de tratamento e de comparação que sejam também associadas aos resultados de interesse. Devido a esta forte hipótese, os métodos de pareamento são, normalmente, mais úteis quando utilizados em conjunto com um dos outros métodos discutidos.

Os métodos de pareamento utilizam, essencialmente, técnicas estatísticas para construir um grupo de comparação artificial ao identificar, para cada observação possível de tratamento, uma observação de não tratamento (ou um conjunto de observações de não tratamento) que possua a maior quantidade possível de características semelhantes. Considere um caso em que você esteja tentando avaliar o impacto de um programa e exista um conjunto de dados que contenha tanto famílias inscritas no programa quanto as não inscritas - por exemplo, a Pesquisa Demográfica e de Saúde (*Demographic and Health Survey*). O programa que você está tentando avaliar não possui regras claras de alocação (tais como uma seleção aleatória ou um índice de elegibilidade), que expliquem por que algumas famílias se inscreveram no

Conceito-chave:
O método de pareamento usa grandes bases de dados e técnicas estatísticas pesadas para construir o melhor grupo de comparação artificial possível para um dado grupo de tratamento.

programa e outras, não. Em tal contexto, os métodos de pareamento permitirão que você identifique o conjunto de famílias não inscritas que mais se assemelham às famílias de tratamento, com base nas características disponíveis em sua base de dados. Essas famílias não inscritas "pareadas" serão o grupo de comparação que você usará para estimar o cenário contrafactual.

Encontrar um bom par para cada participante do programa requer aproximar o máximo possível as variáveis ou *determinantes* que expliquem a decisão das pessoas de participar do programa. Infelizmente, isso é mais fácil dizer do que fazer. Se a lista de características relevantes observadas for muito grande, ou se cada característica assume muitos valores, poderá ser difícil identificar um par correspondente para cada uma das unidades do grupo de tratamento. À medida que você aumenta o número de características ou dimensões com as quais quer parear as unidades que se inscreveram no programa, você pode incorrer na chamada "maldição da dimensionalidade". Por exemplo, se você usar somente três características importantes para identificar o grupo de comparação correspondente - como idade, sexo e local de nascimento - você provavelmente encontrará pares dentre os não inscritos para todos os inscritos em um programa; entretanto, haverá o risco de deixar de fora outras características potencialmente importantes. No entanto, se você aumentar a lista de variáveis, digamos, com a inclusão da quantidade de crianças, do número de anos de estudo, da idade da mãe, da idade do pai e assim por diante, a sua base de dados poderá não conter um bom par para a maioria dos participantes do programa, a menos que contenha uma quantidade muito grande de observações. A Figura 7.1 ilustra o pareamento baseado em quatro características: idade, sexo, meses de desemprego e diploma do Ensino Médio.

Figura 7.1 Pareamento Exato com Quatro Características

Unidades tratadas			
Idade	Sexo	Meses desempregado	Diploma Secundário
9	1	3	0
35	1	12	1
41	0	17	1
23	1	6	0
55	0	21	1
27	0	4	1
24	1	8	1
46	0	3	0
33	0	12	1
40	1	2	0

Unidades tratadas			
Idade	Sexo	Meses desempregado	Diploma Secundário
24	1	8	1
38	0	2	0
58	1	7	1
21	0	2	1
34	1	20	0
41	0	17	1
46	0	9	0
41	0	11	1
19	1	3	0
27	0	4	0

Fonte: Autores, resultado de várias fontes.

Felizmente, a "maldição da dimensionalidade" pode ser facilmente resolvida através de um método chamado "pareamento por escore de propensão" (Rosenbaum e Rubin, 1983). Nessa abordagem, não será mais necessário tentar fazer o pareamento entre cada unidade inscrita com cada unidade não inscrita, que tenha exatamente o mesmo valor para todas as características de controle observadas. Ao invés disso, para cada unidade no grupo de tratamento e no conjunto de não inscritos, calcula-se a probabilidade de uma unidade vir a se inscrever no programa, com base nos valores observados das suas características, o chamado escore de propensão. Esse escore é um único número, variando de 0 a 1, que resume todas as características observadas das unidades que influenciam a probabilidade de inscrição no programa.

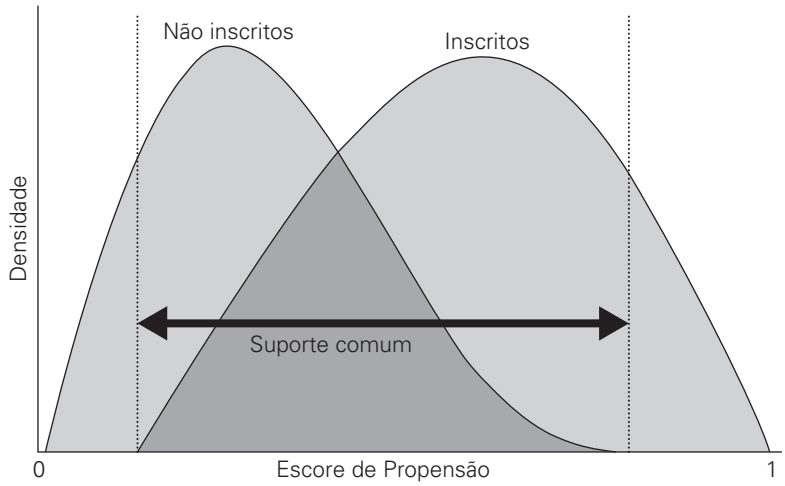
Uma vez que o escore de propensão tenha sido calculado para todas as unidades, poderá então ser feito o pareamento entre as unidades no grupo de tratamento e as unidades no conjunto de não inscritos que tenham escores de propensão o mais próximos possível.[1] Essas "unidades próximas" serão o grupo de comparação e são usadas para produzir uma estimativa do contrafactual. O método de pareamento por escore de propensão tenta simular a alocação aleatória para os grupos de tratamento e comparação, ao escolher unidades para o grupo de comparação que possuam escores de propensão semelhantes às unidades no grupo de tratamento. Considerando que o pareamento por escore de propensão não é um método de alocação aleatória, embora tente simular tal método, ele pertence à categoria dos métodos ditos quase-experimentais.

A diferença nos resultados (Y) entre as unidades de tratamento ou inscritas e suas unidades de comparação correspondentes produz o impacto estimado do programa. Em resumo, o impacto do programa é estimado ao se compararem os resultados médios de um grupo de tratamento ou de inscritos ao resultado médio de um grupo de unidades estatisticamente correspondentes, com o pareamento baseado nas características observadas nos dados disponíveis.

Para que o pareamento por escore de propensão gere estimativas válidas do impacto de um programa, todas as unidades de tratamento precisam ser pareadas satisfatoriamente com uma unidade de não inscritos.[2] Pode acontecer que, para algumas unidades inscritas, nenhuma unidade no conjunto de não inscritos tenha escores de propensão semelhantes. Em termos técnicos, pode haver uma "falta de suporte comum", ou seja, ausência de sobreposição entre os escores de propensão do grupo de tratamento ou grupo de inscritos e os do conjunto dos não inscritos.

A figura 7.2 fornece um exemplo da falta de suporte comum. A probabilidade de cada unidade na amostra se inscrever no programa, é

Figura 7.2 Pareamento por Escore de Propensão e Suporte Comum



Fonte: autores com base em várias fontes.

primeiramente estimada com base nas características observadas de cada unidade. Com base nisso, a cada unidade é atribuído um escore de propensão - em outras palavras, a probabilidade estimada da participação da unidade no programa. A figura mostra a distribuição dos escores de propensão separadamente, para inscritos e não inscritos.

Fundamentalmente, essas distribuições não se sobrepõem perfeitamente. No meio da distribuição, os pares são relativamente fáceis de encontrar porque os inscritos e os não inscritos possuem características semelhantes. No entanto, as unidades com escores de propensão previstos próximos a 1 não podem ser pareadas às unidades não inscritas com escores de propensão semelhantes. Intuitivamente, as unidades que têm alta probabilidade de se inscrever no programa são tão distintas das unidades não inscritas que não podemos encontrar um bom par para elas. Uma falta de suporte comum, portanto, aparece nas extremidades, ou caudas, da distribuição dos escores de propensão.

Jalan e Ravallion (2003a) resumem os passos a serem seguidos ao se aplicar o pareamento por escore de propensão.[3] Primeiramente, você precisa de pesquisas representativas e altamente comparáveis para que seja possível identificar unidades que se inscreveram e unidades que não se inscreveram no programa. Em segundo lugar, você precisa reunir as duas amostras e estimar a probabilidade de que cada indivíduo se inscreva no

programa, com base nas características individuais observadas na pesquisa. Esse passo gera os escores de propensão. Em terceiro lugar, você restringe a amostra às unidades para as quais aparece o suporte comum na distribuição do escore de propensão. Em quarto lugar, para cada unidade inscrita, você localiza um subgrupo de unidades não inscritas que possuem escores de propensão semelhantes. Em quinto lugar, você compara os resultados das unidades de tratamento, ou inscritas, aos resultados de seus pares de comparação, ou unidades não inscritas. A diferença no resultado médio entre uma unidade de tratamento de observação e o subgrupo de unidades pareadas é a medida do impacto que pode ser atribuído ao programa para aquela unidade em particular. Em sexto lugar, a média destes impactos individuais gera o efeito médio de tratamento estimado.

De forma geral, é importante lembrar-se de dois problemas fundamentais sobre o método de pareamento. Primeiro, o pareamento deve ser feito usando-se características de linha de base. Segundo, o método de pareamento é tão bom quanto as características usadas no pareamento, de modo que é fundamental haver um grande número de características básicas.

Utilizando Técnicas de Pareamento para Selecionar Famílias Participantes e Não-Participantes no Programa de Subsídio ao Seguro Saúde

Tendo aprendido sobre técnicas de pareamento, você agora pode questionar se seria possível melhorar as estimativas prévias do impacto do Programa Subsídio ao Seguro Saúde (HISP). Você decide usar algumas técnicas de pareamento para selecionar um grupo de famílias inscritas e não inscritas, que pareçam ser semelhantes com base em um conjunto de características observadas. Primeiro você estima a probabilidade de uma unidade vir a se inscrever no programa, com base nos valores observados de suas características (as "variáveis explicativas"), tais como a idade do chefe da família e do cônjuge, seu nível de educação, se o chefe de família é mulher, se a família é indígena e assim por diante. Conforme apresentado na tabela 7.1, a probabilidade de que uma família se inscreva no programa é menor se a família for mais velha, tiver um nível educacional mais alto ou possuir um banheiro ou uma grande extensão de terras. Por outro lado, ser indígena, ter família mais numerosa e ter chão de terra batida aumenta a probabilidade de que uma família se inscreva no programa. Então, de modo geral, parece que as famílias de menor renda e de menor nível educacional

têm maior probabilidade de se inscrever, o que é uma boa notícia para um programa que visa atingir famílias pobres.

Agora que você já estimou a probabilidade de que cada família se inscreva no programa (o escore de propensão), o passo seguinte é restringir a amostra àquelas famílias nos grupos de inscritos e não inscritos para as quais você pode encontrar um par no outro grupo. Para cada família inscrita, você localiza um subgrupo de famílias não inscritas que possuam escores de propensão semelhantes. A tabela 7.2 compara os resultados médios das famílias inscritas e de seus pares de comparação dentre as famílias não inscritas.

Para obter o impacto estimado usando o método de pareamento, é necessário, primeiramente, calcular o impacto para cada família tratada individualmente (utilizando as famílias de comparação correspondentes de cada família tratada) e, então, calcular a média destes impactos individuais. A tabela 7.3 mostra que o impacto estimado a partir desse procedimento é de uma redução de $8,30 nos gastos das famílias com saúde.

Tabela 7.1 Estimando o Escore de Propensão com base nas Características Observadas

Variável dependente: *Inscritos* = 1	
Variáveis explicativas / características	**Coeficiente**
Idade do chefe de família (anos)	−0,022**
Idade do cônjuge (anos)	−0,017**
Educação do chefe de família (anos)	−0,059**
Educação do cônjuge (anos)	−0,030**
Chefe de família é mulher = 1	−0,067
Indígena = 1	0,345**
Número de membros no domicílio	0,216**
Chão de terra = 1	0,676**
Banheiro = 1	−0,197**
Hectares de terra	−0,042**
Distância ao hospital (km)	0,001*
Constante	0,664**

Fonte: Autores.

Observação: regressão *probit*. A variável dependente é 1 se a família se inscreveu no HISP e 0 se ela não se inscreveu. O coeficiente representa a contribuição de cada variável explicativa / característica listada à probabilidade de que uma família se inscreva no HISP.

* Significativo a 5%, ** Significativo a 1%.

Tabela 7.2 Caso 7 — Impacto do HISP Usando Pareamento (Comparação de Médias)

	Inscritos	Comparação	Diferença	Teste t
Gastos das famílias com saúde	7,8	16,1	–8,3	–13,1

Fonte: Autores.

Tabela 7.3 Caso 7 — Impacto do HISP Usando Pareamento (Análise de Regressão)

	Regressão Linear Multivariada
Impacto estimado nos gastos das famílias com saúde	–8,3** (0,63)

Fonte: Autores.

Observação: Os erros padrão se encontram entre parênteses.

*** Significativo a 1%.*

PERGUNTA 7

A. Quais são as suposições básicas requeridas para se aceitar o resultado do caso 7?

B. Compare o resultado do caso 7 ao resultado do caso 3. Por que você acha que os resultados são tão diferentes?

C. Com base nos resultados do caso 7, o HISP deveria ser estendido nacionalmente?

O Método de Pareamento em Ação

Embora a técnica de pareamento requeira uma quantidade expressiva de dados e tenha outras limitações estatísticas, é um método relativamente versátil e que tem sido usado para avaliar programas de desenvolvimento em diversos contextos. Dois casos ilustrativos são detalhados nos Quadros 7.1 e 7.2

Limitações do Método de Pareamento

Embora os procedimentos de pareamento possam ser aplicados em muitos contextos, independentemente das regras de alocação do programa, eles possuem várias limitações.

Primeiro eles exigem um conjunto extenso de dados e grandes amostras que, mesmo quando disponíveis, não impedem a falta de suporte comum entre o grupo de tratamento - ou de inscritos - e o conjunto de não participantes. Em segundo lugar, o pareamento somente pode ser executado com base em características observadas; por definição, não podemos incorporar características não observadas no cálculo do escore de propensão. Sendo assim, para que o procedimento de pareamento identifique um grupo de

Quadro 7.1: Programa de Ajuda ao Emprego e Renda na Argentina

Jalan e Ravallion (2003a) usaram técnicas de pareamento por escore de propensão para avaliar o impacto do programa argentino de ajuda ao emprego "A Trabajar" sobre a renda. Em resposta à crise macroeconômica de 1996–97 na Argentina, o governo introduziu o *A Trabajar* rapidamente, sem usar qualquer técnica de seleção aleatória ou coletar qualquer dado de linha de base. Por essas razões, os pesquisadores escolheram usar as técnicas de pareamento para avaliar o impacto do programa. Nesse contexto, usar as técnicas de pareamento também possibilita analisar como o ganho de renda varia entre as famílias ao longo da distribuição de renda pré-intervenção.

Em meados de 1997, uma pesquisa foi realizada tanto com os participantes quanto com os não participantes. Para estimar o impacto do programa por pareamento de escore de propensão, Jalan e Ravallion consideraram um grande conjunto de cerca de 200 características básicas (tanto em nível familiar quanto comunitário) que foram medidas na pesquisa. Por exemplo, a estimativa da equação de escore de propensão mostrou que os participantes do programa eram mais pobres e tinham maior probabilidade de serem casados, de serem homens chefes de família e de participarem ativamente de associações de moradores.

Após computar as estimativas dos escores de propensão, os autores restringiram sua análise a unidades cujos escores de propensão caíram na área de suporte comum, onde os escores de propensão de participantes e não participantes se sobrepõem. Ao fazer o pareamento entre os participantes e seus vizinhos mais próximos dentre os não participantes na área de suporte comum, e ao calcular a média das diferenças de renda entre todos esses grupos pareados, eles estimaram que o programa resultou em um ganho de renda médio equivalente a cerca de metade do salário pago pelo programa de ajuda ao emprego. Os pesquisadores verificaram a robustez dos resultados para vários procedimentos de pareamento. Eles destacam que suas estimativas poderiam estar enviesadas devido a características não observadas. Na verdade, ao usar os métodos de pareamento nunca podemos descartar totalmente a hipótese de viés causado por variáveis não observadas - esta é a limitação mais séria desses métodos.

Fonte: Jalan e Ravallion 2003a.

Quadro 7.2: Água Encanada e Saúde Infantil na Índia

Jalan e Ravallion (2003b) usaram os métodos de pareamento para averiguar os efeitos de se ter água encanada em casa sobre a prevalência e duração de diarreia entre as crianças com menos de 5 anos de idade na zona rural da Índia. Em particular, os pesquisadores avaliaram uma política para expandir o acesso à água encanada para entender como os ganhos podem variar em função de circunstâncias familiares, tais como renda e nível educacional. Esse impacto é difícil de ser detectado porque também pode depender de iniciativas de cuidados com a saúde providos privadamente pelos pais e que também afetam a incidência de diarreia, tais como ferver a água, fornecer boa nutrição ou usar sais de hidratação oral quando a criança está doente.

Os pesquisadores usaram dados de uma grande pesquisa realizada em 1993–94 pelo Conselho Nacional de Pesquisa Econômica Aplicada (*The National Council of Applied Economic Research*), que continha informações sobre o estado de saúde e educação de 33.000 famílias rurais de 16 estados da Índia. Este rico grupo de dados permitiu aos pesquisadores usarem o pareamento por escore de propensão tanto em nível individual quanto de município, equilibrando os grupos de tratamento e de comparação segundo a probabilidade prevista de receberem água encanada durante a campanha nacional.

A avaliação descobriu que o acesso à água encanada reduziu a diarreia - sua prevalência seria 21% maior e duraria 29% a mais sem a água encanada. No entanto, estes impactos não são observados em grupos de baixa renda a menos que as mulheres da família possuam nível educacional superior à educação fundamental. De fato, Jalan e Ravallion descobriram que os impactos da água encanada na saúde são maiores e mais significativos em famílias em que mulher possui melhor educação. Eles concluíram que o estudo ilustra a necessidade de se combinar investimentos em infraestrutura, tais como água encanada, com outros programas para melhorar a educação e reduzir a pobreza.

Fonte: Jalan e Ravallion 2003b.

comparação válido, precisamos ter certeza de que não existe nenhuma diferença sistemática em características não observadas entre as unidades de tratamento e de comparação correspondentes[4] que possa influenciar o resultado *(Y)*. Uma vez que não podemos *provar* que não há características não observadas que afetem tanto a participação quanto os resultados, temos que *supor* que elas não existem. Este é, geralmente, um pressuposto muito forte. Embora o pareamento ajude a controlar as características básicas *observadas*, jamais poderemos descartar o viés resultante de características *não observadas*. Em resumo, a suposição de que não houve viés de seleção

oriundo de características não observadas é muito forte e, o que é mais problemático ainda, não pode ser testado.

Em geral, os métodos de pareamento são menos robustos do que os outros métodos de avaliação que discutimos anteriormente. Por exemplo, os métodos de seleção aleatória não requerem hipóteses não testáveis de que não existem variáveis não observadas que expliquem tanto a participação no programa quanto os resultados. Eles também não requerem grandes amostras e grande número de características básicas, como o pareamento de escore de propensão.

Na prática, os métodos de pareamento são tipicamente usados quando as opções de seleção aleatória, regressão descontínua e *diferença-em-diferenças* não são possíveis. Muitos autores usam o chamado pareamento *ex-post*, quando nenhum dado de linha de base está disponível para o resultado de interesse ou para as características de contexto. Eles usam uma pesquisa realizada após o início do programa (isto é, *ex-post*) para inferir quais eram as características básicas das pessoas na linha de base (por exemplo, idade, estado civil) e, então, fazem o pareamento entre o grupo de tratamento e o grupo de comparação usando essas características inferidas. Naturalmente, isso é arriscado: eles podem inadvertidamente fazer o pareamento com base em características que também foram afetadas pelo programa e, neste caso, o resultado da estimativa seria inválido ou enviesado.

Por outro lado, quando dados de linha de base estão disponíveis, o método de pareamento com base nas características de contexto pode ser muito útil, se combinado a outras técnicas - por exemplo, a *diferença-em-diferenças*, que leva em conta heterogeneidades não observadas e constantes no tempo. O pareamento também é mais útil quando a regra de alocação do programa é conhecida, caso em que o pareamento pode ser executado com base naquela regra (vide capítulo 8).

Até aqui, deve ter ficado claro para os leitores que as avaliações de impacto são melhor desenhadas antes que um programa comece a ser implementado. Uma vez que o programa se inicia, se não houver uma forma de influenciar como ele é alocado e não houver dados de linha de base coletados, as opções válidas e disponíveis para a avaliação serão poucas ou inexistentes.

Notas

1. Na prática, muitas definições do que constitui o escore de propensão "mais próximo" são usadas para executar o pareamento. Os controles mais próximos podem ser definidos com base em uma estratificação dos escores de

propensão - a identificação dos vizinhos mais próximos da unidade de tratamento, baseado na distância, dentro de um determinado raio - ou com o uso de técnicas de *kernel*. Considera-se boa prática verificar a robustez dos resultados de pareamento, com o uso de vários algoritmos de pareamento.

2. A discussão sobre métodos de pareamento neste livro enfoca o pareamento um-a-um. Vários outros tipos de pareamento, tais como o pareamento múltiplo (de um para vários) ou de substituição/não substituição não serão discutidos. Em todo caso, a estrutura conceitual descrita aqui ainda seria aplicável.
3. Rosenbaum 2002 apresenta uma análise detalhada dos métodos de pareamento.
4. Para leitores com experiência em econometria, isto significa que a participação é independente dos resultados, dadas as características básicas usadas para se fazer o pareamento.

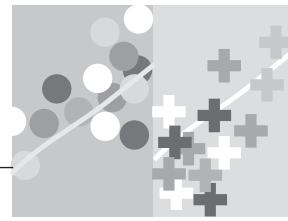
Referências

Jalan, Jyotsna, e Martin Ravallion. 2003a. "Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching." *Journal of Business & Economic Statistics* 21 (1): 19–30.

——. 2003b. "Does Piped Water Reduce Diarrhea for Children in Rural India?" *Journal of Econometrics* 112 (1): 153–73.

Rosenbaum, Paul. 2002. *Observational Studies*. 2.ª edição Springer Series in Statistics. Nova York: Springer-Verlag.

Rosenbaum, Paul, e Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies of Causal Effects." *Biometrika* 70 (1): 41–55.



CAPÍTULO 8

A Combinação de Métodos

Vimos que a maioria dos métodos de avaliação de impacto somente produzem estimativas válidas do contrafactual sob certas suposições. O maior risco na aplicação de qualquer método é as suposições subjacentes serem falsas, resultando em estimativas enviesadas do impacto do programa. Esta seção avalia estes problemas metodológicos e discute estratégias para reduzir o risco de viés. Uma vez que o risco de viés resulta principalmente de desvios das suposições subjacentes, enfocaremos a forma de verificar tais suposições.

Em alguns métodos de avaliação, pode ser verificada a validade das suposições das quais cada um depende. Em outros métodos, não se pode verificar a validade de modo direto, mas é possível usar vários testes, conhecidos como "testes de falseamento", para melhorar a confiança nos pressupostos subjacentes do método escolhido. Os testes de falseamento são parecidos com os testes de resistência. Se falharem, é um forte sinal de que os pressupostos por trás do método não se sustentam, neste contexto específico. No entanto, se não falharem, isto significa apenas um suporte experimental aos pressupostos: você nunca poderá estar completamente certo de que sejam verdadeiros. O quadro 8.1 apresenta uma lista de testes de verificação e falseamento usados para avaliar se um método é apropriado no contexto da sua avaliação. A lista contém questões práticas que podem ser respondidas pela análise dos dados de linha de base.

Quadro 8.1: Checklist para Testes de Verificação e Falseamento

Alocação aleatória

A alocação aleatória é o método mais robusto para estimar os cenários contrafatuais; é considerado o padrão de ouro da avaliação de impacto. Alguns testes básicos ainda devem ser considerados, para avaliar a validade desta estratégia de avaliação em um determinado contexto:

- As características de linha de base estão balanceadas? Compare as características de linha de base do grupo de tratamento e do grupo de comparação.[a]
- Houve algum descumprimento das regras de alocação? Verifique se todas as unidades elegíveis receberam o tratamento ou se nenhuma unidade inelegível recebeu o tratamento. Caso apareça algum descumprimento, use o método de oferta aleatória.
- A quantidade de unidades nos grupos de tratamento e de comparação é suficientemente grande? Caso não seja, você pode querer combinar a alocação aleatória com o método de *diferença-em-diferenças*.

Oferta aleatória

Descumprimento na alocação aleatória equivale à oferta aleatória.

- As características de linha de base estão equilibradas? Compare as características de linha de base das unidades às quais o programa está sendo oferecido e aquelas às quais o programa não está sendo oferecido.

Promoção aleatória

A promoção aleatória leva a estimativas válidas do cenário contrafactual se a campanha promocional aumenta substancialmente a participação no programa, sem afetar diretamente os resultados de interesse.

- As características de linha de base estão equilibradas entre as unidades que recebem a campanha promocional e aquelas que não a recebem? Compare as características de linha de base dos dois grupos.
- A campanha promocional afeta substancialmente a participação no programa? Deveria. Compare as taxas de participação no programa nas amostras promovidas e não promovidas.
- A campanha promocional afeta diretamente os resultados? Não deveria. Isto não pode ser diretamente testado, então precisamos nos basear em teoria e no bom senso para nos orientarmos.

Regressão descontínua (RDD)

A regressão descontínua requer que o índice de elegibilidade seja contínuo em torno do ponto de corte e que as unidades sejam comparáveis ao redor do ponto de corte.

- O índice é contínuo em torno do ponto de corte no instante de tempo da linha de base?
- Há algum descumprimento do corte para o tratamento? Teste se todas as unidades elegíveis e nenhuma inelegível receberam o tratamento. Caso você encontre algum descumprimento, será preciso combinar o RDD com técnicas mais avançadas para corrigir esta "descontinuidade nebulosa".[b]

Diferenças-em-diferenças (DD)

O método diferença-em-diferenças assume que as tendências nos resultados sejam semelhantes nos grupos de comparação e de tratamento antes da intervenção e que os únicos fatores que explicam as mudanças

(continuação)

Quadro 8.1 *continuação*

nos resultados entre os dois grupos são constantes no tempo.

- Os resultados se moveriam em paralelo nos grupos de tratamento e de comparação na ausência do programa? Isto pode ser avaliado com o uso de vários testes de falseamento, como: (1) os resultados nos grupos de tratamento e comparação andavam em paralelo antes da intervenção? Se duas rodadas de dados estavam disponíveis antes do início do programa, teste para ver se aparece qualquer diferença entre os dois grupos. (2) E as variáveis de resultado falsas, que não deveriam ser afetadas pelo programa? Elas se movem em paralelo antes e depois do início da intervenção nos grupos de tratamento e de comparação?
- Execute a análise diferença-em-diferenças usando vários grupos de comparação plausíveis. Você obteve estimativas semelhantes do impacto do programa?
- Execute a análise diferença-em-diferenças usando os seus grupos de tratamento e de comparação escolhidos e uma variável de resultado falsa que não deveria ser afetada pelo programa. Você deve encontrar um impacto zero do programa nesta variável de resultado.
- Execute a análise diferença-em-diferenças usando a sua variável de resultado escolhida com dois grupos que você sabe que não foram afetados pelo programa. Você deve encontrar impacto zero do programa.

Pareamento

O pareamento se baseia na suposição de que as unidades inscritas e não inscritas são semelhantes em termos de quaisquer variáveis não observadas que poderiam afetar tanto a probabilidade de participação no programa quanto o resultado (Y).

- A participação no programa é determinada por variáveis que não podem ser observadas? Isto não pode ser testado diretamente, então precisamos agir com base na teoria e no bom senso.
- As características observadas estão bem equilibradas entre os subgrupos pareados? Compare as características observadas de cada unidade de tratamento e seu grupo correspondente de comparação.
- Pode ser encontrada uma unidade de comparação para cada unidade de tratamento? Verifique se existe suporte comum suficiente na distribuição dos escores de propensão. Pequenas áreas de suporte comum indicam que os indivíduos inscritos e não inscritos são muito diferentes e isso lança dúvida sobre a confiabilidade do método de pareamento.

Fonte: Autores.

a. Conforme citado anteriormente, por razões estatísticas, nem todas as características observadas têm que ser semelhantes nos grupos de tratamento e de comparação para que a aleatorização tenha bons resultados. Mesmo quando as características dos dois grupos são verdadeiramente iguais, pode-se esperar que 5% das características apresentarão uma diferença estatisticamente significante quando usamos um nível de confiança de 95% para o teste.

b. Embora não elaboraremos esta técnica aqui, os leitores podem desejar saber que se combinaria o RDD a uma abordagem de variável instrumental. Usar-se-ia a localização à esquerda ou à direita do ponto de corte como uma variável instrumental para a participação real do programa na primeira etapa de uma estimativa dos mínimos quadrados em dois estágios.

Combinação de Métodos

Embora todos os métodos de avaliação corram o risco de viés, o risco pode, às vezes, ser reduzido com o uso de uma combinação de métodos. Ao combinar os métodos, podemos contrabalançar as limitações de um único método e, consequentemente, aumentar a robustez do cenário contrafactual estimado.

O método de diferença-em-diferenças com pareamento por escore de propensão (DD *matching*) é um exemplo de combinação de métodos. Conforme discutido anteriormente, o pareamento por escore de propensão simples não pode responder por características não observadas que possam explicar por que um grupo escolhe se inscrever em um programa e que também pode afetar os resultados. Por outro lado, o pareamento combinado com diferença-em-diferenças cuida, pelo menos, de qualquer característica não observada que seja constante no decorrer do tempo entre os dois grupos. É executado da seguinte forma:

- Inicialmente, execute o pareamento baseando-se nas características de linha de base observadas (conforme discutido no capítulo 7).
- Em seguida, aplique o método diferença-em-diferenças para estimar o contrafactual da mudança nos resultados em cada subgrupo de unidades pareadas.
- Por último, calcule a média destas diferenças duplas nos subgrupos pareados.

O quadro 8.2 fornece um exemplo de uma avaliação que utilizou o método de diferença-em-diferenças com pareamento por escore de propensão, na prática.

O método de regressão descontínua com diferenças-em-diferenças (DD RDD) é um segundo exemplo de métodos combinados. Há que se lembrar que o RDD assume que as unidades em ambos os lados do limiar de elegibilidade sejam muito semelhantes. Na medida em que algumas diferenças permanecem entre as unidades de cada um dos lados do limiar, adicionar o método de *diferença-em-diferenças* nos permite controlar as diferenças nas características não observadas que não variam ao longo do tempo. Pode-se executar o DD RDD extraindo-se a diferença dupla dos resultados para as unidades em ambos os lados do corte de elegibilidade.

Cumprimento Imperfeito

O cumprimento imperfeito (ou conformidade imperfeita) é uma discrepância entre o estado de tratamento pretendido e o estado de tratamento real.

Quadro 8.2: *Diferença-em-Diferenças* com Pareamento
Chão de cimento, saúde infantil, e felicidade das mães no México

O programa Piso Firme, no México, oferece aos domicílios com chãos de terra até 50 metros quadrados de chão de concreto. O Piso Firme começou como um programa local no estado de Coahuila, sendo então adotado nacionalmente. Cattaneo et al. (2009) aproveitaram a variação geográfica para avaliar o impacto deste esforço de larga escala de melhoramento de residências nos resultados de saúde e bem-estar.

Os pesquisadores usaram o método *diferença-em-diferenças,* em conjunto com o pareamento, para comparar os domicílios em Coahuila a domicílios semelhantes no estado vizinho de Durango que, à época da pesquisa, ainda não havia implementado o programa. Para melhorar a comparabilidade entre os grupos de tratamento e de comparação, os pesquisadores limitaram sua amostra a famílias em cidades vizinhas que vivem próximas da fronteira entre os dois estados. Eles retiraram as amostras dos quarteirões das duas cidades que tinham as características pré-intervenção mais semelhantes, segundo o censo de 2002.

Usando a oferta do chão de cimento como uma variável instrumental para ter, de fato, chão de cimento, os pesquisadores recuperaram o efeito do tratamento sobre os tratados do estimador ITT e descobriram que o programa levou a uma redução de 18,2% na presença de parasitas, uma redução de 12,4% na prevalência de diarreia e a uma redução de 19,4% na prevalência de anemia. Além disso, eles puderam usar a variabilidade na área total de chão de fato coberto por cimento para prever que uma substituição completa dos chãos de terra por cimento em um domicílio levaria a uma redução de 78% na infestação de parasitas, uma redução de 49% na diarreia, uma redução de 81% na anemia e uma melhoria de entre 36% e 96% no desenvolvimento cognitivo. Os autores também coletaram dados sobre o bem-estar dos adultos e descobriram que o chão de cimento deixa as mães mais felizes, com um aumento de 59% na satisfação com o domicílio, 69% a mais na satisfação com a qualidade de vida, redução de 52% em uma escala de avaliação da depressão e uma redução de 45% na escala de avaliação de estresse percebido.

Cattaneo et al. (2009) concluíram ilustrando que o Piso Firme tem um impacto absoluto maior no desenvolvimento cognitivo infantil por um custo mais baixo do que o programa de larga escala de transferência condicionada de renda mexicano, o Oportunidades/Progresa, assim como programas comparáveis de suplemento nutricional e estimulação cognitiva na primeira infância. O chão de cimento também previne mais infecções parasitais do que o tratamento antiparasitário comum. Os autores declaram que os programas para substituir o chão de terra por chão de cimento têm maior probabilidade de melhorar a saúde infantil em termos de custo-eficiência, em contextos semelhantes.

Fonte: Cattaneo et al. 2009.

Discutimos isto ao fazer referência à alocação aleatória mas, na realidade, o cumprimento imperfeito é um potencial problema encontrado na maioria dos métodos de avaliação de impacto. Antes de você poder interpretar as estimativas de impacto produzidas por qualquer método, você precisa saber se ocorreu um cumprimento imperfeito no programa.

O cumprimento imperfeito possui duas manifestações: (1) algumas unidades planejadas para receber o tratamento podem não recebê-lo e (2) algumas unidades planejadas para fazer parte do grupo de comparação podem receber o tratamento. O cumprimento imperfeito pode ocorrer em uma variedade de formas:

- Nem todos os participantes planejados do programa participam do programa de fato. Às vezes, unidades às quais o programa é oferecido optam por não participar.
- O programa não é oferecido a alguns participantes planejados, devido a erros administrativos ou de execução.
- O programa é oferecido equivocadamente a algumas unidades do grupo de comparação e estas se inscrevem no programa.
- Algumas unidades do grupo de comparação tentam participar do programa mesmo quando este não é oferecido a elas. Às vezes, isto é chamado de "contaminação" do grupo de comparação. Se a contaminação afetar uma grande porção do grupo de comparação, estimativas não enviesadas do contrafactual não podem ser obtidas.
- O programa é designado aos beneficiários com base em um escore contínuo de priorização, mas o corte de elegibilidade não é executado rigorosamente.
- Migração seletiva ocorre com base no estado de tratamento. Podemos, por exemplo, usar o método diferença-em-diferenças para comparar os resultados de municípios tratados e não tratados, mas as pessoas podem escolher ir para outro município se não gostarem do estado de tratamento de seu próprio município.

Em geral, na presença de cumprimento imperfeito, os métodos padrão de avaliação de impacto produzem estimativas da intenção de tratar. No entanto, as estimativas de tratamento sobre os tratados podem ser recuperadas a partir das estimativas da intenção de tratar, usando a abordagem de variável instrumental.

No capítulo 4, apresentamos a intuição básica para lidar com o cumprimento imperfeito no contexto da alocação aleatória. Utilizando-se um ajuste

para a porcentagem dos que cumprem com as regras de participação no programa na amostra de avaliação, pode-se descobrir o impacto do tratamento sobre as unidades tratadas, a partir da estimativa da intenção de tratar. Este "ajuste" pode ser estendido a outros métodos, através da aplicação da abordagem de variável instrumental mais geral. A variável instrumental ajuda a resolver ou corrigir o viés que pode resultar do cumprimento imperfeito. No caso da oferta aleatória, usamos uma variável 0/1 (ou "*dummy*"), que assume o valor 1 se a unidade foi originalmente alocada ao grupo de tratamento e 0 se a unidade foi originalmente alocada ao grupo de comparação. Durante a fase de análise, a variável instrumental é, geralmente, utilizada no contexto de uma *regressão em dois estágios,* que permite identificar o impacto do tratamento naqueles que cumprem as regras de participação do programa.

A lógica da abordagem da variável instrumental pode ser estendida ao contexto de outros métodos de avaliação:

- No contexto da regressão descontínua, a variável instrumental a ser usada é uma variável que assume os valores 0 ou 1, indicando se uma unidade está localizada do lado inelegível ou do lado elegível do ponto de corte.
- No contexto do método *diferença-em-diferenças* e da migração seletiva, uma possível variável instrumental para a localização do indivíduo após o início do programa seria a localização do indivíduo antes do anúncio do programa.

Apesar da possibilidade de "ajustar" o cumprimento imperfeito usando variáveis instrumentais, é importante lembrar-se de dois pontos:

1. Do ponto de vista técnico, não é desejável ter-se uma grande parcela do grupo de comparação inscrita no programa. Os avaliadores e formuladores de políticas envolvidos na avaliação de impacto devem trabalhar juntos para manter esta fração em um nível mínimo.
2. O método da variável instrumental somente é válido sob determinadas circunstâncias – certamente não é uma solução universal.

Efeitos de Transbordamento

Mesmo quando o grupo de comparação não é diretamente beneficiado pelo programa, ele pode ser indiretamente afetado pelos transbordamentos do grupo de tratamento. Um exemplo interessante é discutido por Kremer e Miguel (2004), que examinaram o impacto de dar vermífugos a crianças de escolas no Quênia (quadro 8.3). Os vermes intestinais são parasitas que

Quadro 8.3: Trabalhando com Efeitos de Transbordamentos

Vermifugação, externalidades e educação no Quênia

O projeto de vermifugação na escola primária, em Busia, Quênia, foi executado pela ONG holandesa Child Support Africa, em cooperação com o Ministério da Saúde, e foi elaborado para testar uma variedade de aspectos do tratamento e prevenção de vermes. O projeto envolveu 75 escolas, com uma participação total de mais de 30.000 estudantes entre 6 e 18 anos de idade. As escolas foram tratadas com vermífugos de acordo com as recomendações da Organização Mundial de Saúde e também receberam educação preventiva contra vermes, na forma de palestras de saúde, cartazes e treinamento para os professores.

Devido a restrições administrativas e financeiras, a implementação foi feita em fases, por ordem alfabética, com o primeiro grupo de 25 escolas começando em 1998, o segundo grupo em 1999 e o terceiro grupo em 2001. Ao selecionar as escolas aleatoriamente, Kremer e Miguel (2004) puderam estimar o impacto da vermifugação em uma escola e identificar transbordamentos nas demais escolas usando variações exógenas nas escolas de controle mais próximas das escolas de tratamento. Embora a conformidade com o método aleatório tenha sido relativamente alta (com 75% dos designados ao tratamento recebendo vermífugos e somente uma pequena porcentagem das unidades do grupo de comparação recebendo o tratamento), os pesquisadores também puderam aproveitar o descumprimento para determinar as externalidades interescolares de saúde, ou transbordamentos. Kremer e Miguel (2004) descobriram que o efeito da externalidade interescolar foi a redução de 12 pontos percentuais na proporção de infecções parasitárias, desde as moderadas até as severas, enquanto o efeito adicional direto de se tomar realmente a medicação foi de 14 pontos percentuais a mais. Também em termos de externalidades entre as escolas, a cada mil alunos adicionais que frequentam uma escola de tratamento associa-se uma redução de 26 pontos percentuais nas infecções moderadas e severas. Estes efeitos na saúde também levaram a um aumento na participação escolar de, pelo menos, sete pontos percentuais e reduziram o absenteísmo em um quarto, pelo menos. Não foi encontrado nenhum impacto significativo nos testes de aprendizagem.

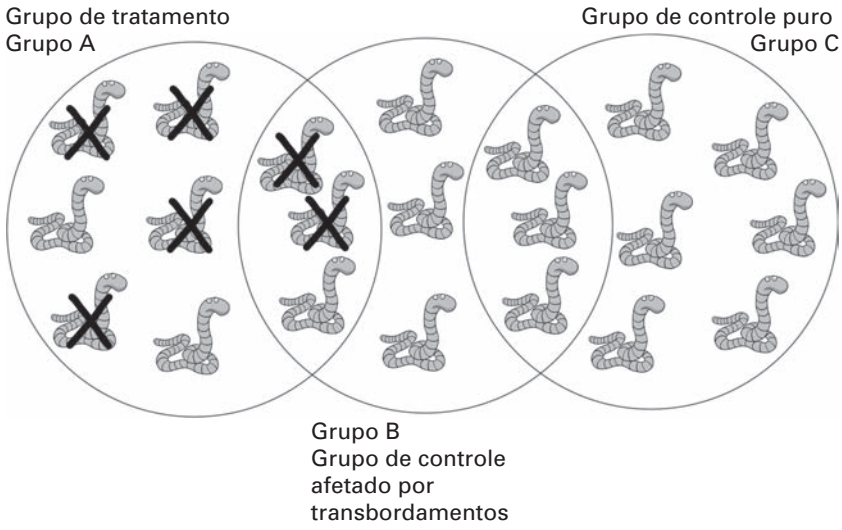
Devido ao baixíssimo custo do tratamento contra vermes e aos efeitos educacionais relativamente altos, os pesquisadores concluíram que a vermifugação é uma forma eficiente de melhorar as taxas de participação nas escolas. O estudo também ilustra que as doenças tropicais como os vermes podem desempenhar um papel importante nos resultados educacionais e fortalecer as alegações de que a alta carga de doenças na África estaria contribuindo para o baixo nível de renda. Desse modo, os autores argumentam que o estudo reforça a ideia de subsídios públicos para tratamento de doenças que tenham efeitos de transbordamento similares nos países em desenvolvimento.

Fonte: Kremer e Miguel 2004.

podem ser transmitidos de uma pessoa para outra por meio do contato com matéria fecal contaminada. Quando uma criança recebe vermífugos, sua "carga parasitária" diminuirá. Mas também diminuirá a carga parasitária das pessoas que vivem no mesmo ambiente, já que elas não estarão mais em contato com os vermes da criança. No exemplo queniano, portanto, quando o vermífugo foi dado às crianças em uma escola, ele beneficiou não somente aquelas crianças (benefício direto), mas também as crianças das escolas circunvizinhas (benefício indireto).

Conforme retratado na figura 8.1, a vermifugação no grupo de escolas A também diminui a quantidade de vermes que afetam as escolas do grupo B, que não estavam no programa mas que estão localizadas perto das escolas do grupo A. No entanto, as escolas não participantes do programa e mais distantes das escolas do grupo A – as chamadas escolas do grupo C - não experimentam os efeitos de tais transbordamentos porque o vermífugo usado com o grupo A não mata nenhum dos vermes que afetam o grupo C. Kremer e Miguel (2004) descobriram que a vermifugação reduziu significativamente o absenteísmo escolar, não somente nas escolas participantes do programa (ao comparar o grupo A com o grupo C), mas também nas escolas próximas que não foram incluídas no programa (ao comparar o grupo B ao grupo C).

Figura 8.1 Efeitos de transbordamento



Fonte: Autores.

Como transbordamentos podem ocorrer, é importante que o avaliador verifique se estes não afetam todo o grupo de comparação. Contanto que haja unidades de comparação em número suficiente que não sejam afetadas por transbordamentos (o grupo C no exemplo da vermifugação), será possível estimar o impacto do programa ao comparar os resultados das unidades de tratamento aos resultados das unidades de comparação "puras". A desvantagem é que a avaliação não poderá generalizar os efeitos do tratamento estimado para toda a população. Se, na fase de desenho, você espera que o programa tenha efeitos de transbordamento, é possível adaptar o desenho da avaliação para produzir resultados melhores. Primeiro, o desenho precisa ser capaz de identificar um grupo de comparação puro. Será, então, possível generalizar o impacto do programa estimado. Em segundo lugar, o desenho também deve possibilitar a estimativa da magnitude dos efeitos de transbordamento, ao identificar um grupo de comparação que tenha probabilidade de ser afetado por transbordamento. De fato, geralmente os efeitos de transbordamento são de interesse, pois constituem impactos indiretos do programa.

A figura 8.1 ilustra como é possível estimar tanto o impacto de um programa quando qualquer efeito de transbordamento. O grupo A recebe o medicamento. O efeito do medicamento repercute no grupo B. O grupo C está mais afastado e, portanto, não recebe o efeito de transbordamento da medicação. Este desenho pode ser obtido ao se atribuir o tratamento aleatoriamente entre duas unidades próximas e uma unidade similar distante. Nesta estrutura simples, o impacto do programa pode ser estimado ao comparar os resultados do grupo A com os resultados do grupo C; os efeitos de transbordamento podem ser estimados comparando os resultados do grupo B aos do grupo C.

Considerações Adicionais

Somando-se ao cumprimento imperfeito e aos transbordamentos, outros fatores também precisam ser considerados quando elabora-se uma avaliação de impacto. Estes fatores são comuns à maioria das metodologias que discutimos até agora e tendem a ser mais difíceis de mitigar.[1]

Ao planejar uma avaliação, deve-se determinar o momento certo da coleta de dados. Se um programa levar tempo para gerar impactos nos resultados, a coleta prematura de dados acabará sem conseguir detectar qualquer impacto do programa (vide, por exemplo, King e Behrman 2009). De outro modo, se a pesquisa de seguimento for feita tardiamente, não será possível captar os efeitos do programa a tempo de informar os formuladores de

políticas. Nos casos onde se deseja estimar o impacto do programa tanto de curto quanto de longo prazo, várias rodadas de pós-intervenção ou dados de seguimento precisarão ser coletados. O capítulo 10 apresentará outras orientações em relação aos melhores marcos temporais para a execução da avaliação.

Se o impacto de um programa estiver sendo estimado em todo um grupo, os resultados podem mascarar algumas diferenças nas respostas ao tratamento dentre os diferentes beneficiários. A maioria dos métodos de avaliação de impacto assume que um programa afeta os resultados de um modo simples e linear para todas as unidades da população. No entanto, podem surgir problemas quando o tamanho da resposta depender de forma não linear do tamanho da intervenção, ou quando um grupo com uma alta intensidade de tratamento for comparado com um grupo com baixa intensidade de tratamento. Se você achar que diferentes subpopulações possam ter experimentado o impacto de um programa de maneira muito diferente, então você pode considerar ter diferentes amostras para cada subpopulação. Digamos, por exemplo, que você esteja interessado em conhecer o impacto de um programa de merenda escolar sobre o desempenho escolar de meninas, mas somente 10% dos estudantes são meninas. Neste caso, até mesmo uma "grande" amostra aleatória de estudantes pode não conter um número suficiente de meninas para permitir-lhe estimar o impacto do programa nas meninas. Para o seu desenho amostral da avaliação, você deveria, então, estratificar a amostra por sexo e incluir um número suficientemente alto de meninas para permitir a detecção de um determinado tamanho de efeito.

Ao realizar uma avaliação de impacto, você pode induzir respostas comportamentais indesejáveis da população que estiver estudando, que podem vir a limitar a validade externa dos resultados da avaliação. Por exemplo, o "efeito Hawthorne" ocorre quando o mero fato de você observar as unidades faz com que elas se comportem de maneira diferente (Levitt e List 2009). O "efeito John Henry" acontece quando unidades de comparação trabalham mais arduamente para compensar a falta do tratamento. A antecipação pode levar a outro tipo de efeito comportamental não desejável. Em um programa com expansão aleatória, as unidades no grupo de comparação podem esperar receber o programa no futuro e começar a mudar o seu comportamento antes que o programa realmente chegue. Se você tiver motivos para acreditar que estas respostas comportamentais não planejadas possam existir, então construir grupos de comparação adicionais, que não sejam afetados pela intervenção, é, às vezes, uma opção que, de fato, permite testar explicitamente tais respostas.

Um Plano Alternativo para a Avaliação

Às vezes, mesmo com o melhor método de avaliação de impacto e as melhores intenções, as coisas não acontecem exatamente conforme o planejado. Em uma recente experiência de um programa de treinamento para o trabalho, o organismo de implementação planejou selecionar aleatoriamente os participantes do conjunto de candidatos, acreditando que haveria excesso de candidatos ao programa. Devido ao alto índice de desemprego na população-alvo, foi antecipado que o conjunto de candidatos para o programa de capacitação para o trabalho seria muito maior do que o número de vagas disponíveis. Infelizmente, a campanha para o programa não foi tão efetiva quanto se esperava, e por fim, o número de candidatos estava um pouco abaixo do número de vagas disponíveis para o treinamento. Sem o excesso de candidatos de onde retirar um grupo de comparação e sem um plano alternativo, a tentativa inicial de avaliar o programa teve que ser completamente esquecida. Este tipo de situação é comum, já que há mudanças não previstas no contexto operacional ou político de um programa. Portanto, é útil ter um plano alternativo, caso a primeira escolha de metodologia não funcione. A parte 3 deste livro discute mais detalhadamente os aspectos operacionais e políticos da avaliação.

Planejar usar vários métodos de avaliação de impacto também é uma boa prática do ponto de vista metodológico. Caso tenha dúvidas se um de seus métodos estará enviesado, é possível comparar os resultados com outros métodos. Quando o programa é implementado com etapas de expansão aleatórias (vide capítulo 10), o grupo de comparação será finalmente incorporado ao programa. Isto limita o tempo durante o qual o grupo de comparação está disponível para a avaliação. Se, entretanto, somando-se ao método de alocação aleatória, um método de promoção aleatória também for implementado, então um grupo de comparação estará disponível durante todo o período do programa. Antes da incorporação do último grupo ao programa, existirão dois grupos alternativos de comparação (da alocação aleatória e da promoção aleatória) embora, no longo prazo, somente o grupo de comparação de promoção aleatória permanecerá.

Nota

1. No capítulo 3 são discutidas outras fontes de limites à validade externa relacionadas a vieses no processo amostral e resultantes de diferentes atrições nos grupos de tratamento e de comparação.

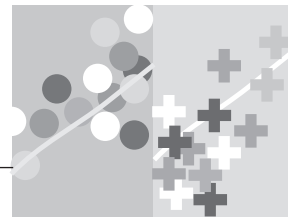
Referências

Cattaneo, M., Galiani, S., Gertler, P., Martinez, S. & Titiunik, R. (2009). Housing, Health e Happiness. *American Economic Journal: Economic Policy* 1 (1): 75–105.

King, E. M. & Behrman, J. (2009). Timing and Duration of Exposure in Evaluations of Social Programs. *Observatório de Pesquisa do Banco Mundial* 24 (1): 55–82.

Kremer, M. & Miguel, E. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica* 72 (1): 159–217.

Levitt, S. & List, J. (2009). Was There Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. *NBER Working Paper 15016*. Cambridge, MA: National Bureau of Economic Research.



CAPÍTULO 9

Avaliando programas multifacetados

Até agora, temos discutido programas que incluem somente um tipo de tratamento. Na realidade, muitas perguntas relevantes sobre políticas surgem no contexto de programas multifacetados - isto é, programas que combinam várias opções de tratamento.[1] Os formuladores de políticas podem ter interesse em saber não somente se um programa funciona, mas também se o programa funciona melhor do que o outro ou a um custo mais baixo. Por exemplo, se quisermos aumentar a frequência escolar, é mais efetivo implementar intervenções do lado da demanda (tais como transferências de renda às famílias) ou intervenções do lado da oferta (tais como maiores incentivos para os professores)? Se introduzirmos as duas intervenções juntas, elas funcionarão melhor do que se estivessem isoladas? Em outras palavras, elas se complementam? Alternativamente, se a relação custo-efetividade de um programa é uma prioridade, você pode se perguntar qual é o nível ótimo de serviços que o programa deveria entregar. Por exemplo, qual é a duração ótima de um programa de capacitação profissional? Um programa de 6 meses tem um efeito maior na taxa de inserção no mercado de trabalho de seus capacitados do que um programa de 3 meses? Em caso afirmativo, a diferença é grande o suficiente para justificar os recursos adicionais necessários para um programa de 6 meses?

Além de simplesmente estimar o impacto de uma intervenção sobre um resultado de interesse, as avaliações de impacto podem ajudar a responder a questões mais amplas, tais como:

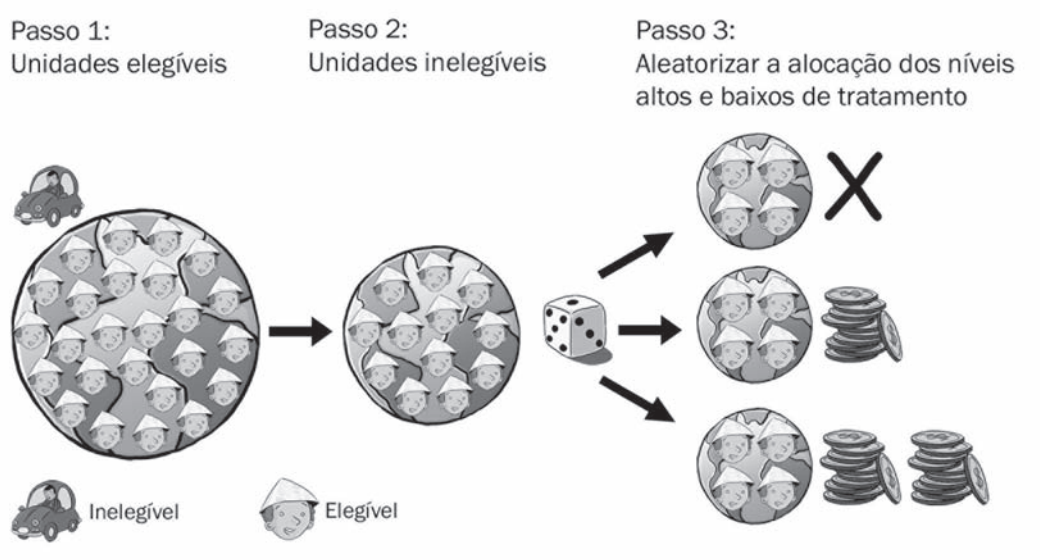
- *Qual é o impacto de um tratamento comparado com o de outro tratamento?* Por exemplo, qual é o impacto no desenvolvimento cognitivo das crianças de um programa que fornece capacitação aos pais, comparado a uma intervenção nutricional?
- *O impacto conjunto de um primeiro tratamento e um segundo tratamento é maior do que a soma dos dois impactos individuais?* Por exemplo, o impacto total de uma intervenção junto aos pais e de uma intervenção nutricional é maior, menor ou igual à soma dos efeitos das duas intervenções individualmente?
- *Qual é o impacto adicional de um tratamento de intensidade maior, comparado a um tratamento de intensidade menor?* Por exemplo, qual é o efeito no desenvolvimento cognitivo das crianças desnutridas se um assistente social visita o domicílio a cada quinze dias, em comparação a uma visita mensal?

Este capítulo oferece exemplos de como elaborar avaliações de impacto para dois tipos de programa multifacetados: programas com vários níveis do mesmo tratamento e programas com múltiplos tratamentos. Iremos primeiramente discutir como desenhar uma avaliação de impacto para um programa com vários níveis de serviço, e, então, trataremos de como desmembrar os vários tipos de impacto de um programa com múltiplos tratamentos. A discussão assume que estejamos usando o mecanismo de alocação aleatória, mas pode ser generalizada a outros métodos.

Avaliando Programas com Diferentes Níveis de Tratamento

É relativamente fácil desenhar uma avaliação de impacto para um programa com vários níveis de tratamento. Suponha uma tentativa de avaliação do impacto de um programa que apresente dois níveis de tratamento: alto (por exemplo, visitas quinzenais) e baixo (visitas mensais), onde se pretenda avaliar o impacto das duas opções e saber quanto as visitas adicionais afetarão os resultados. Para fazer isso, pode-se realizar um sorteio para decidir quem recebe o nível alto de tratamento, quem recebe o nível baixo de tratamento e quem é alocado para o grupo de comparação. A Figura 9.1 ilustra este processo.

Figura 9.1 Etapas na Alocação Aleatória de dois Níveis de Tratamento



Fonte: Autores.

Assim como em uma alocação aleatória padrão, o primeiro passo é definir a população de unidades elegíveis para o programa. O segundo passo é selecionar uma amostra aleatória de unidades a serem incluídas na avaliação, a chamada amostra de avaliação. Na terceira etapa, já de posse da amostra de avaliação, selecionam-se unidades aleatoriamente para o grupo que receberá o nível alto de tratamento, para o grupo que receberá o nível baixo de tratamento ou para o grupo de comparação. Serão criados três grupos distintos resultantes da seleção aleatória para níveis de tratamento múltiplos:

- O grupo A constitui o grupo de comparação.
- O grupo B recebe o nível baixo de tratamento.
- O grupo C recebe o nível alto de tratamento.

Quando executada corretamente, a seleção aleatória garante que os três grupos sejam semelhantes. Pode-se, portanto, estimar o impacto do tratamento de nível alto pela comparação do resultado médio do grupo C com o resultado médio do grupo A. Da mesma forma, é possível estimarmos o impacto do tratamento de nível baixo pela comparação do resultado médio do grupo B com o do grupo A. Finalmente, é possível avaliar se o tratamento de nível alto tem um impacto maior do que o tratamento de nível baixo, através da comparação dos resultados médios dos grupos B e C.

A estimativa do impacto de um programa com mais de dois níveis de tratamento segue a mesma lógica. Se houver três níveis de tratamento, o processo de aleatoriedade criará três grupos de tratamento diferentes, mais um grupo de comparação. Em geral, com *n d*iferentes níveis de tratamento, haverá *n* grupos de tratamento, mais um grupo de comparação.

Quando a alocação aleatória não for praticável, outros métodos de avaliação devem ser utilizados. Felizmente, todos os métodos de avaliação descritos até agora são capazes de analisar o impacto relativo de diferentes níveis de tratamento. Suponha, por exemplo, que você esteja interessado em avaliar o impacto da diferentes quantias de dinheiro oferecidos a alunos em um programa de bolsas que visa aumentar as matrículas nas escolas do Ensino Médio. Uma bolsa de 60 dólares é concedida aos 25 alunos com as melhores notas em cada escola ao final do ensino fundamental e um bolsa de 45 dólares é concedida ao grupo de alunos com as seguintes 25 notas mais altas. Os alunos com as piores classificações não recebem bolsas. Neste contexto, pode ser usado um método de regressão descontínua para comparar as notas dos testes dos alunos, não somente em torno do limiar de 45 dólares, mas também perto do limite de 60 dólares. Filmer e Schady (2009) apresentam os resultados desta avaliação no Camboja, na qual não encontraram evidência de que a bolsa de 60 dólares tenha aumentado o nível de matrículas em maior grau do que a bolsa de 45 dólares.

Avaliando Múltiplos Tratamentos com Desenhos Cruzados

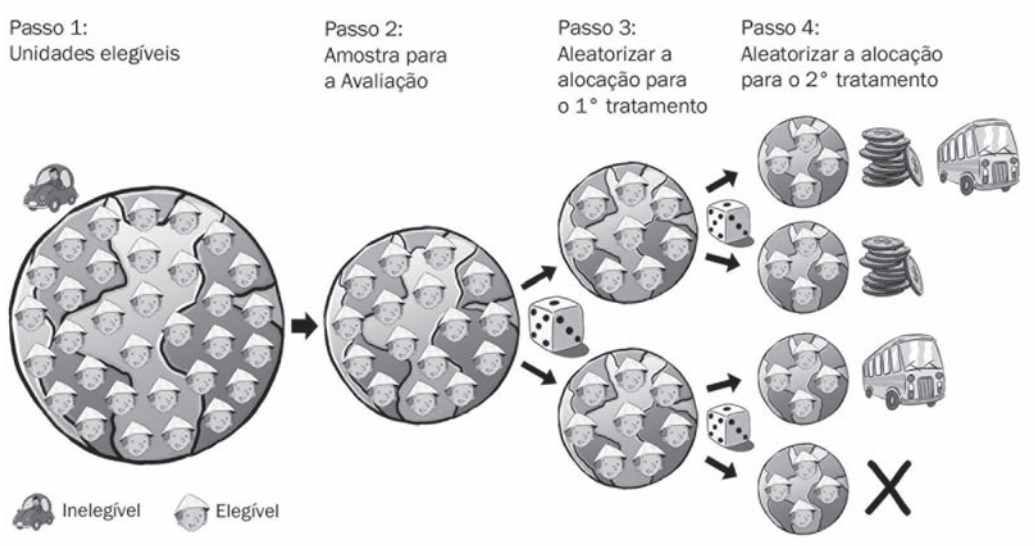
Além de comparar os vários níveis de tratamento, você pode querer comparar opções de tratamento completamente diferentes. De fato, os formuladores de políticas preferem, geralmente, comparar os méritos relativos de diferentes intervenções, ao invés de conhecer o impacto de uma única intervenção.

Imagine que você queira avaliar o impacto sobre a matrícula escolar de um programa com duas intervenções diferentes: transferência condicionada de renda para as famílias dos alunos e transporte de ônibus gratuito para a escola. Você pode querer conhecer o impacto de cada intervenção separadamente; pode também querer saber se a combinação das duas é melhor do que somente a soma dos efeitos individuais. Visto do ponto de vista dos participantes, o programa está disponível em três formas diferentes: somente a transferência condicionada de renda, somente transporte

gratuito de ônibus ou uma combinação de transferência condicionada de renda e transporte de ônibus gratuito.

A seleção aleatória para um programa com duas intervenções é muito parecida com o processo para um programa com uma única intervenção. A principal diferença é a necessidade de realizar vários sorteios independentes, ao invés de um só. Isto produz um método cruzado (crossover), às vezes também chamado de método transversal cruzado (cross-cutting). A Figura 9.2 ilustra este processo. Como antes, o passo 1 define a população de unidades elegíveis para o programa. O passo 2 seleciona uma amostra aleatória de unidades elegíveis da população para formar a amostra de avaliação. Uma vez obtida a amostra de avaliação, o terceiro passo é alocar aleatoriamente unidades da amostra de avaliação em um grupo de tratamento e um grupo de controle. Na quarta etapa, usa-se um segundo sorteio para alocar aleatoriamente um subconjunto do grupo de tratamento para receber a segunda intervenção. Finalmente, na quinta etapa, é realizado outro sorteio para alocar um subconjunto do grupo de controle inicial para que receba a segunda intervenção, enquanto o outro subconjunto permanece como um controle "puro".

Figura 9.2 Etapas na Alocação Aleatória de duas Intervenções



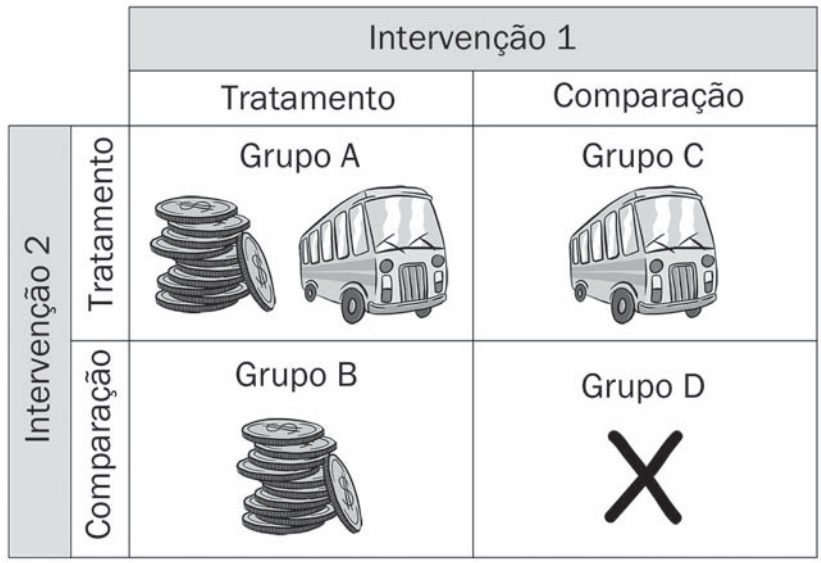
Fonte: Autores.

Como resultado da alocação aleatória nos dois tratamentos, você terá criado quatro grupos, conforme ilustrado na Figura 9.2

- O grupo A recebe ambas as intervenções (transferências de renda e transporte de ônibus).
- O grupo B recebe a intervenção 1, mas não a intervenção 2 (somente transferência de renda).
- O grupo C não recebe a intervenção 2, mas recebe a intervenção 1 (somente transporte de ônibus).
- O grupo D não recebe nem a intervenção 1 nem a intervenção 2; ele constitui o grupo de comparação puro.

Quando implementada corretamente, a alocação aleatória garante que os quatro grupos sejam semelhantes. Você pode, portanto, estimar o impacto da primeira intervenção ao comparar o resultado do grupo B com resultado do grupo de comparação puro, o grupo D. Você também pode estimar o impacto da segunda intervenção comparando o resultado do grupo C com

Figura 9.3 Grupos de Tratamento e de Comparação em um Programa com duas Intervenções



Fonte: Autores.

o resultado do grupo de comparação puro. Além disso, o método também possibilita comparar o impacto incremental de receber a segunda intervenção quando uma unidade já recebe a primeira. Comparar os resultados do grupo A com os do grupo B resultará no impacto da segunda intervenção para as unidades que já recebem a primeira intervenção e comparar os resultados do grupo A aos do grupo C resultará no impacto da primeira intervenção para as unidades que já recebem a segunda intervenção.

A descrição acima usou o exemplo da alocação aleatória para explicar como uma avaliação de impacto pode ser concebida para um programa com duas intervenções diferentes. Quando um programa envolver mais do que duas intervenções, pode-se aumentar a quantidade de sorteios e continuar subdividindo a amostra de avaliação para compor grupos que receberão as várias combinações de intervenções. Desenhos com tratamentos múltiplos e múltiplos níveis de tratamento também podem ser implementados. Mesmo se o número de grupos aumentar, a teoria básica por trás do desenho da avaliação permanecerá a mesma, como descrita anteriormente.

No entanto, avaliar mais do que uma ou duas intervenções criará desafios práticos, tanto para a avaliação quanto para as operações do programa, na medida em que a complexidade do método aumentará exponencialmente com a quantidade dos ramos de tratamento. Para avaliar uma intervenção, somente dois grupos são necessários: um grupo de tratamento e um grupo de comparação. Para avaliar duas intervenções, são necessários quatro grupos: três grupos de tratamento e um grupo de comparação. Se você fosse avaliar três intervenções, incluindo todas as combinações possíveis entre as três intervenções, seriam necessários 2 x 2 x 2 = 8 grupos na avaliação. Em geral, para que todas as combinações possíveis entre as n intervenções fossem incluídas na avaliação, seriam necessários 2^n grupos. Além disso, para podermos fazer a distinção entre as diferenças nos resultados entre os diferentes grupos, cada grupo deve conter uma quantidade suficiente de unidades de observação para garantir suficiente poder estatístico. De fato, detectar as diferenças entre diferentes ramos de intervenção pode requerer amostras maiores do que quando o tratamento é comparado a um controle puro. Se os dois ramos de tratamento tiverem bons resultados no sentido de ocasionar mudanças nos resultados desejados, serão necessárias amostras maiores para detectar diferenças potencialmente menores entre os dois grupos.

Finalmente, métodos transversais também podem ser usados em desenhos de avaliações que combinam vários métodos de avaliação (Quadros 9.1 e 9.2). As regras operacionais que orientam a alocação de cada tratamento determinarão que combinação de métodos deve ser usada. Por exemplo, pode ser que o primeiro tratamento seja alocado com base em uma

Quadro 9.1: Testando Alternativas de Programa de Prevenção do Vírus HIV/AIDS no Quênia

Duflo et al. (2006) usaram um método transversal cruzado para avaliar o impacto de uma série de programas de prevenção ao vírus HIV/AIDS em dois distritos rurais do oeste queniano. O estudo se baseou em uma amostra de 328 escolas, que foram divididas em seis grupos conforme apresentado na tabela abaixo, que resume o desenho do programa. Foi alocada aleatoriamente a cada grupo uma combinação diferente de três tratamentos. Os tratamentos incluíam: o oferecimento de um programa de capacitação de professores, visando melhorar a capacidade de ensinar a parte do currículo nacional de educação referente ao vírus HIV/AIDS; estimular os alunos a realizar debates sobre o papel dos preservativos e concursos de redação sobre prevenção; redução dos custos da educação com a fornecimento de uniformes escolares gratuitos aos alunos (vide tabela).

Resumo do Método do Programa

Grupo	Número de escolas	Programa nacional	Capacitação de professores reforço	Debate sobre preservativos e redação (primavera 2005)	Reduzir o custo da educação (primavera 2003 e outono 2004)
1	88	Sim			
2	41	Sim	Sim		
3	42	Sim	Sim	Sim	
4	83				Sim
5	40	Sim	Sim		Sim
6	40	Sim	Sim	Sim	Sim

Os pesquisadores constataram que, após dois anos, o programa de capacitação de professores tinha tido pouco impacto no conhecimento dos alunos, na atividade sexual reportada, no uso de preservativos ou na gravidez na adolescência, embora tenha melhorado o ensino do currículo nacional. Os debates e o concurso de redação aumentaram o conhecimento autoreportado e o uso de preservativo, sem aumentar a atividade sexual autoreportada. Finalmente, reduzir o custo da educação ao fornecer uniformes escolares reduziu tanto a taxa de evasão quanto a gravidez na adolescência. Portanto, os pesquisadores concluíram que fornecer uniformes escolares provou ter melhores resultados na redução da gravidez na adolescência do que a capacitação de professores em relação ao currículo nacional sobre o vírus HIV/AIDS.

Fonte: Duflo et al. 2006.

Quadro 9.2: Testando Alternativas de Programas para Monitorar a Corrupção na Indonésia

Na Indonésia, Olken (2007) usou um método transversal cruzado inovador para testar diversos métodos de controle da corrupção, desde uma estratégia de coerção de cima para baixo até a supervisão comunitária. Ele usou uma metodologia de alocação aleatória em mais de 600 povoados que estavam construindo estradas como parte de um projeto de melhoria da infraestrutura nacional.

Um dos tratamentos múltiplos incluía a seleção aleatória de alguns povoados para informar-lhes que seu projeto de construção seria auditado por um fiscal do governo. Então, para testar a participação da comunidade na supervisão, os pesquisadores implementaram duas intervenções. Eles distribuíram convites para encontros comunitários de prestação de contas e forneceram formulários de comentários, que poderiam ser enviados anonimamente. Para medir os níveis de corrupção, uma equipe independente de engenheiros e pesquisadores coletou amostras básicas das novas estradas, estimou o custo dos materiais usados e, então, comparou seus cálculos aos orçamentos informados.

Olken descobriu que aumentar as auditorias governamentais (com o aumento da chance de ser auditado de 4% para 100%) reduziu o desperdício de despesas em 8 pontos percentuais (de 24%). Aumentar a participação da comunidade na supervisão teve um impacto no desperdício de trabalho, mas não no desperdício de recursos financeiros. Os formulários de comentários foram efetivos somente quando distribuídos para crianças na escola, a fim de que entregassem às suas famílias e não quando distribuídos pelos líderes comunitários.

Fonte: Olken 2007.

pontuação de elegibilidade, mas o segundo seja alocado de forma aleatória. Neste caso, o desenho pode usar uma regressão descontínua na primeira intervenção e um método de alocação aleatória na segunda intervenção.

Nota

1. Vide Banerjee e Duflo (2009) para uma discussão mais aprofundada.

Referências

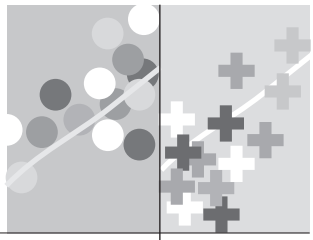
Banerjee, A. & Duflo, E. (2009). The Experimental Approach to Development Economics. *NBER Working Paper 14467.* Cambridge, MA: National Bureau of Economic Research.

Duflo, E., Dupas, P., Kremer, M. & Sinei, S. (2006). Education and HIV/AIDS Prevention: Evidence from a Randomized Evaluation in Western Kenya. *Policy Research Working Paper 402*. Washington, DC: Banco Mundial.

Filmer, D. & Schady, N. (2009). School Enrollment, Selection and Test Scores. *Policy Research Working Paper 4998*. Washington, DC: Banco Mundial.

Olken, B. (2007). Monitoring Corruption: Evidence from a Field

Experiment in Indonesia. *Journal of Political Economy* 115 (2): 200–49.



Parte 3

COMO IMPLEMENTAR UMA AVALIAÇÃO DE IMPACTO

Na parte 1 deste livro, discutimos por que uma avaliação de impacto deveria ser realizada e quando vale a pena fazê-la. Em princípio, as avaliações devem ser concebidas para tratar de questões que precisam ser respondidas para fins de formulação de políticas, como, por exemplo, em negociações de orçamento ou decisões sobre a possibilidade de expandir um programa de nutrição, aumentar benefícios de bolsas de estudo ou executar uma reforma em um hospital. Os objetivos e questões da avaliação devem fluir diretamente das questões sobre as políticas. Uma vez que esteja claro quais políticas precisam ser avaliadas e quais questões a avaliação deve abordar, será necessário desenvolver uma teoria de mudança, como uma cadeia de resultados para o programa, que permitirá escolher indicadores apropriados. Na parte 2 deste livro, descrevemos uma série de métodos para avaliar o impacto dos programas e discutimos suas vantagens e desvantagens, com exemplos para cada método.

Esta terceira parte do livro se concentra nas etapas operacionais da contratação e gestão de uma avaliação de impacto. Essas etapas constituem as bases

deuma avaliação de impacto que responderá às questões de políticas formuladas e estimará o impacto causal do programa. Agrupamos as etapas operacionais de uma avaliação de impacto em quatro grandes fases: *operacionalização do método de avaliação, escolha de uma amostra, coleta de dados e produção e divulgação dos resultados*. A figura da página seguinte ilustra estasequência e os capítulos 10 a 13 tratam de cada uma das quatro fases.

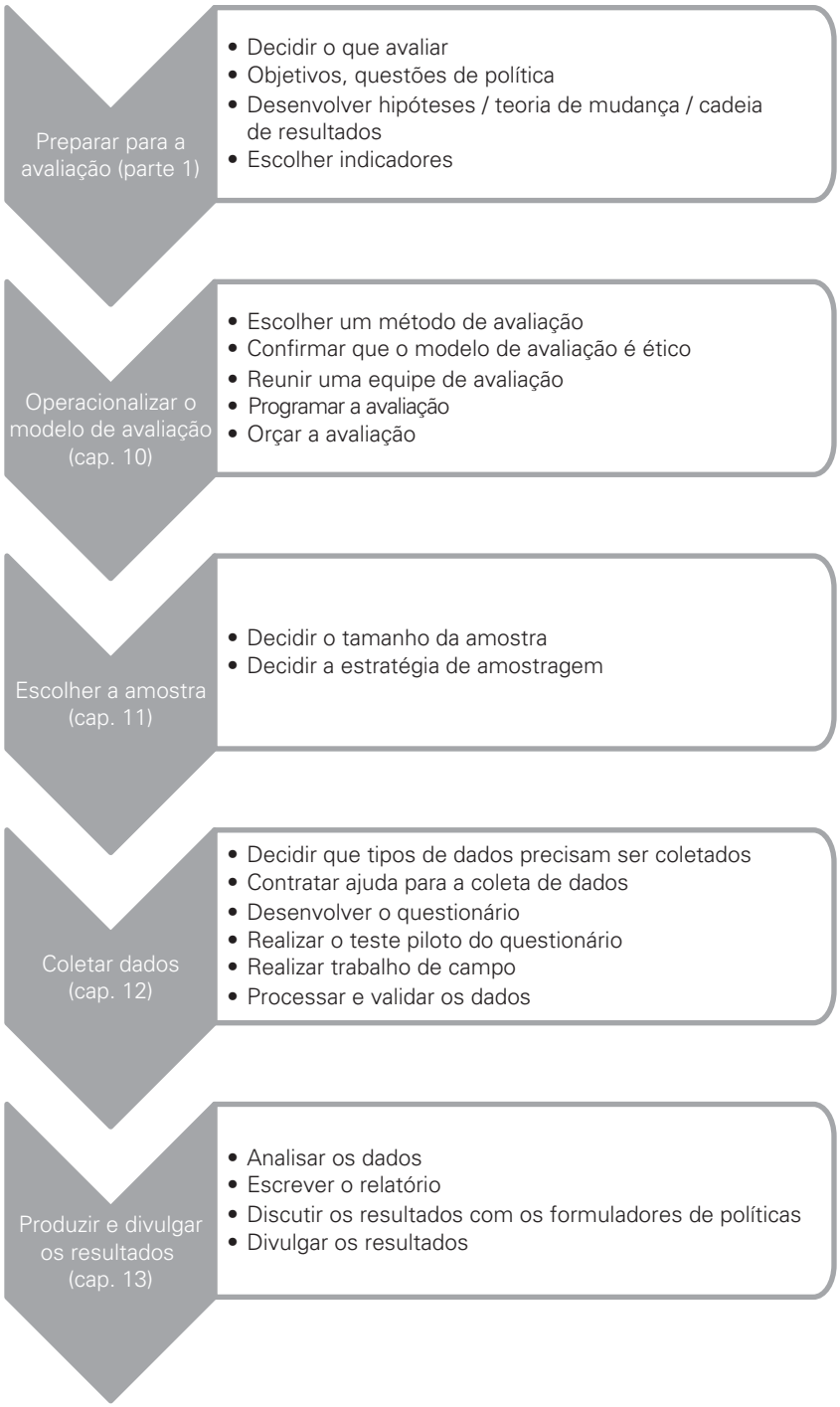
No capítulo 10, discutiremos os principais componentes da operacionalização do método para a avaliação. Isto é, você examinará os planos de implementação do programa e escolherá um desenho de avaliação adequado. Antes que você possa passar à implementação da avaliação, você deve confirmar que o desenho de avaliação proposto é ético. Uma vez que isso esteja claro, você reunirá uma equipe de avaliação, elaborará um orçamento e identificará a fonte de financiamento.

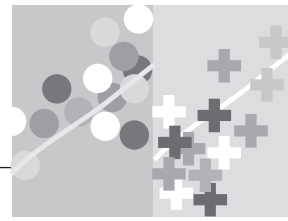
No capítulo 11, discutiremos como sortear entrevistados para as pesquisas e quantos entrevistados são necessários.

No capítulo 12, nós revisaremos os passos da coleta de dados. Tendo em mente as questões de políticas que se almeja responder, bem como o desenho da avaliação, você deve determinar quais dados podem ser extraídos de fontes existentes e decidir que tipos de dados precisam ser coletados. Você deve supervisionar o desenvolvimento de um questionário apropriado para os dados a ser coletados. Uma vez feito isso, você deve contratar uma empresa ou órgão governamental especializado em coleta de dados. Essa entidade recrutará e capacitará o pessoal de campo e realizará o teste piloto do questionário. Depois de realizar os ajustes necessários, a empresa ou agência será capaz de avançar com o trabalho de campo. Finalmente, os dados deverão ser digitalizados ou processados e validados antes que possam ser utilizados.

No capítulo 13, trataremos dos estágios finais da avaliação. Descreveremos os produtos entregues por uma avaliação e o que os relatórios de avaliação devem conter e daremos algumas orientações sobre como divulgar os resultados entre os formuladores de políticas e outras partes interessadas.

Figura P3.1 Roadmap for Implementing an Impact Evaluation





CAPÍTULO 10

Operacionalizando o Desenho de Avaliação de Impacto

Na parte 2, descrevemos diversas metodologias alternativas que produzem grupos de comparação válidos. Com base nesses grupos de comparação, pode ser estimado o impacto causal de um programa. Passamos agora aos aspectos práticos da escolha do método para avaliar o seu programa. Mostraremos que as regras de funcionamento do programa orientarão claramente a forma de geração dos grupos de comparação e, assim, a decisão sobre qual método é mais apropriado para o contexto da política adotada.

Escolhendo um Método de Avaliação de Impacto

A chave para estimar um impacto causal é encontrar um grupo de comparação válido. Na Parte 2, discutimos inúmeros grupos de comparação válidos, incluindo aqueles gerados por alocação aleatória, promoção aleatória, regressão descontínua, diferença-em-diferenças, e pareamento. Neste capítulo, consideraremos a questão de qual método usar em cada situação. O princípio geral é de que as regras de funcionamento do programa guiam a decisão sobre qual método é mais adequado para cada programa e que essas regras podem e devem conduzir o método de avaliação, não o contrário. A avaliação não deve mudar drasticamente os elementos-chave da intervenção em prol de um método de avaliação mais limpo.

Conceito-chave:
As regras de funcionamento do programa determinam qual método de avaliação de impacto pode ser aplicado (não o contrário).

A alocação aleatória é, frequentemente, o método preferido pelos avaliadores. Quando devidamente implementada, gera comparabilidade entre os grupos de tratamento e de comparação em relação a características observadas e não observadas, com baixo risco de viés. Como a alocação aleatória é bastante intuitiva, ela requer uso limitado de econometria e gera um efeito médio de tratamento para a população de interesse. Também torna mais simples e direta a comunicação dos resultados aos formuladores de políticas. No entanto, desenhos aleatórios nem sempre são viáveis, especialmente quando conflitam com as regras de funcionamento do programa.

As regras operacionais mais relevantes para o método de avaliação são aquelas que identificam quem é elegível para o programa e como os indivíduos são selecionados para participar. Grupos de comparação são compostos por indivíduos elegíveis, mas que não podem ser incorporados em determinado momento (por exemplo, quando existe excesso de demanda) ou próximas do limite de participação no programa, de acordo com as regras de focalização ou elegibilidade. É difícil encontrar grupos de comparação válidos se as regras do programa que determinam a elegibilidade e seleção dos beneficiários não forem equitativas, transparentes e responsabilizáveis.[1]

Princípios de Critério de Elegibilidade do Programa

Quase sempre podemos encontrar grupos de comparação válidos, se as regras operacionais de seleção de beneficiários forem equitativas, transparentes e com atribuição clara de responsabilidades (*accountable*):

- Critérios *equitativos* de focalização são regras que classificam ou priorizam a elegibilidade com base em um indicador de necessidade definido de comum acordo, ou sob os quais são oferecidos os benefícios do programa a todos ou onde todos tenham, pelo menos, chances iguais de receber os benefícios.
- Critérios *transparentes* de focalização são regras que são tornadas públicas, de maneira que a sociedade civil possa concordar implicitamente com elas e possa monitorar se estão realmente sendo seguidas. Regras transparentes devem ser quantitativas e facilmente observáveis por terceiros.
- Regras *com atribuição clara de responsabilidade (accountable)* são regras sob a responsabilidade dos técnicos do programa e cuja implementação é a base do desempenho e do reconhecimento pelo trabalho desses técnicos.

Conceito-chave:
Podemos quase sempre encontrar grupos de comparação válidos, se as regras operacionais para a seleção de beneficiários forem equitativas, transparentes, e responsáveis.

Regras equitativas, como discutiremos mais tarde, traduzem-se, na maioria dos casos, em regressões descontínuas ou em alocações aleatórias. Transparência e responsabilização asseguram que os critérios de

focalização sejam verificáveis e realmente sejam aplicados conforme concebidos. Quando as regras operacionais violam esses três princípios de boa governança, enfrentamos desafios, tanto na criação de um programa bem desenhado quanto na realização da avaliação.

As regras operacionais de elegibilidade são transparentes e com atribuição clara quando o governo usa critérios quantificáveis que podem ser verificados externamente e quando torna públicos esses critérios. Esses princípios de boa governança aumentam a probabilidade de que o programa realmente beneficie a população-alvo e são a chave de uma avaliação bem sucedida. Se as regras não forem quantificáveis e verificáveis, a equipe de avaliação terá dificuldade em garantir que a seleção aos grupos de comparação e de tratamento ocorra como desenhada ou, no mínimo, em documentar como ela de fato ocorreu. Se os avaliadores não puderem realmente verificar a seleção, eles não poderão analisar corretamente os dados para calcular os impactos. Compreender as regras de seleção do programa é essencial para identificar o método de avaliação de impacto adequado.

Critérios Operacionais de Elegibilidade do Programa

Regras de funcionamento normalmente regulam quais são os benefícios do programa, como eles são financiados e distribuídos e como o programa seleciona os beneficiários. As regras que regem o financiamento de programas e a inclusão de beneficiários são essenciais para identificar grupos válidos de comparação. As regras que regem a inclusão cobrem a elegibilidade, as regras de alocação no caso de recursos limitados e a incorporação gradual de beneficiários. Mais especificamente, as principais regras que geram um roteiro para os grupos de comparação respondem a três questões operacionais fundamentais relativas a recursos financeiros, elegibilidade e tempo:

1. *Recursos financeiros: o programa tem recursos suficientes para ganhar escala e atingir a cobertura total de todos os beneficiários elegíveis?* Governos e organizações não governamentais nem sempre têm recursos financeiros suficientes para fornecer serviços de programas a todos que sejam elegíveis e que busquem os benefícios. Nesse caso, o governo tem que decidir quais dos candidatos elegíveis receberão os benefícios do programa e quais serão excluídos. Muitas vezes, os programas são limitados a regiões geográficas específicas, áreas rurais ou pequenas comunidades, muito embora possa haver beneficiários elegíveis em outras regiões ou em comunidades maiores.

2. *Critério de focalização do programa: Quem é elegível para os benefícios do programa? O programa é focalizado com base em um corte de elegibilidade,*

ou está disponível a todos? Escolas públicas e cuidados básicos de saúde são, normalmente, oferecidos de maneira universal. Muitos programas utilizam critérios operacionais de focalização do programa que dependem de uma classificação contínua com um ponto de corte. Os programas de pensão, por exemplo, definem um limite de idade, acima do qual os idosos se tornam elegíveis. Programas de transferência de renda frequentemente classificam domicílios com base em seu estado de pobreza estimada e domicílios abaixo de um ponto de corte predeterminado são considerados elegíveis.

3. *Tempo: Como os potenciais beneficiários são inscritos no programa — todos de uma só vez ou em fases graduais?* Muitas vezes, restrições administrativas e de recursos impedem um governo de proporcionar benefícios imediatos a todo seu grupo-alvo. O governo deve implantar o programa ao longo do tempo e, por conseguinte, decidir quem recebe os benefícios primeiro e quem será incorporado mais tarde. Uma abordagem comum é aplicar um programa geograficamente, de forma gradual, ao longo do tempo, incorporando todos os beneficiários elegíveis em um povoado ou região, antes de passar para o próximo.

Identificando e Priorizando Beneficiários

Uma questão operacional crítica embutida em todas as três questões é como os beneficiários são selecionados. Isso, como veremos abaixo, é a chave para identificar grupos de comparação válidos. Os grupos de comparação são naturalmente encontrados entre as populações não elegíveis e mais frequentemente entre as populações que são elegíveis, mas que serão incorporadas mais tarde. A forma como os beneficiários são priorizados depende, em parte, dos objetivos do programa. Trata-se de um programa de pensões para os idosos, de alívio da pobreza focalizado nos pobres ou de imunização disponível a todos?

Para priorizar os beneficiários, o programa deve escolher um indicador que seja quantificável e verificável. Uma vez que seja acordado um indicador de necessidade, o modo como ele será aplicado dependerá, em grande medida, da capacidade do governo de medir e classificar a necessidade. Se o governo puder classificar com precisão os beneficiários com base na necessidade relativa, poderá sentir-se eticamente obrigado a implantar o programa por ordem de necessidade. No entanto, classificar com base na necessidade requer não apenas uma medida quantificável, mas também a capacidade e os recursos para medir esse indicador por indivíduo.

Em alguns casos, a elegibilidade pode ser baseada num indicador contínuo que é barato e fácil de coletar, como a idade para pensões. Por exemplo,

70 anos de idade como ponto de corte para a elegibilidade para pensões é relativamente simples de medir e fácil de aplicar. No entanto, muitas vezes o indicador de elegibilidade não classifica a necessidade relativa dentro da população elegível. Por exemplo, uma pessoa de 69 anos não necessariamente precisa menos de uma pensão do que uma pessoa de 70; uma pessoa de 75 anos de idade não necessariamente precisa da pensão mais do que uma de 72. Nesse caso, o programa pode identificar a população elegível, mas não pode classificar facilmente a necessidade relativa da população elegível.

Outros programas usam critérios de elegibilidade que, em princípio, poderiam ser utilizados tanto para determinar a elegibilidade quanto classificar a necessidade relativa. Por exemplo, muitos programas são direcionados a pessoas pobres, embora os indicadores precisos de pobreza que classificam os domicílios de forma confiável sejam, frequentemente, difíceis de medir e dispendiosos de coletar. Coletar dados de renda ou de consumo de todos os potenciais beneficiários para classificá-los por nível de pobreza é um processo complexo e dispendioso. Em vez disso, muitos programas usam algum tipo de teste baseado em uma análise da renda aproximada dos potenciais beneficiários para estimar os níveis de pobreza. Esses são índices de medidas simples, tais como bens e características sociodemográficas (Grosh et al. 2008). Esses testes de elegibilidade podem padecer de erros de medição, são dispendiosos de implementar e nem sempre permitem a classificação afinada de status socioeconômico ou necessidade, especialmente na base da distribuição de pobreza. Testes de elegibilidade podem ajudar a determinar razoavelmente bem se determinada família está acima ou abaixo de algum ponto de corte bruto, mas podem ser menos precisos ao identificar a distância do ponto de corte. Seu uso permite aos programas identificar os pobres elegíveis, mas não necessariamente classificar a necessidade dentro de uma população elegível.

Em vez de confrontar o custo e a complexidade de classificar os domicílios, muitos programas escolhem classificar em um nível mais alto de agregação, como no nível da comunidade. A hipótese implícita é de que os domicílios dentro das comunidades são basicamente homogêneos, que a grande maioria da população é igualmente elegível e que classificar os domicílios não valeria o custo de identificar e excluir os poucos inelegíveis. Nesse caso, todos dentro de uma comunidade seriam elegíveis para o programa. Embora essa estratégia funcione para pequenas comunidades rurais, ela não funciona mais tão bem à medida que os programas se deslocam para áreas mais urbanizadas, que são mais heterogêneas. A focalização em um nível agregado tem vantagens operacionais óbvias, mas frequentemente não evita a necessidade de classificar beneficiários individuais com base em um indicador objetivo e quantificável de necessidade.

Nos casos em que a agência que financia o programa escolhe não classificar a necessidade porque o processo é muito caro e sujeito a erros, deve usar outros critérios para decidir como determinar a ordem da implementação do programa. Um critério consistente com a boa governança é o da equidade. Uma regra equitativa seria dar a todos os elegíveis a mesma chance de ser os primeiros e selecionar aleatoriamente os potenciais beneficiários para seus lugares na sequência. Esta é uma regra justa e equitativa de alocação, e produz também um desenho de avaliação aleatório, com validade tanto interna quanto externa.

Traduzindo Regras Operacionais em Grupos de Comparação

Na Tabela 10.1, mapeamos os possíveis grupos de comparação para o tipo de programa, com base nas regras operacionais e nas três questões operacionais fundamentais relativas a recursos financeiros, elegibilidade e tempo que formulamos anteriormente. As colunas são divididas dependendo se o programa tem ou não recursos suficientes para cobrir todos os beneficiários potencialmente elegíveis *(recursos financeiros)* e são subdivididas, ainda, entre programas focalizados e universais *(critério de elegibilidade do programa)*. As linhas são divididas em implantação gradual *versus* imediata do programa *(tempo)*. Cada célula lista as potenciais fontes de grupos de comparação válidos. Cada célula é rotulada com um índice em que o primeiro algarismo indica a linha na tabela (A, B) e o segundo indica a coluna (1-4). Por exemplo, a célula A1 refere-se à célula na primeira linha da primeira coluna da tabela. A célula A1 identifica os métodos de avaliação mais adequados para programas com recursos limitados, que são focalizados e implementados gradualmente ao longo do tempo.

A maioria dos programas precisa ser implementada gradualmente ao longo do tempo, quer por restrições financeiras, quer por limitações logísticas ou administrativas. Esse grupo ou categoria cobre a primeira linha da tabela — ou seja, as células A1, A2, A3 e A4. Nesse caso, a regra operacional equitativa, transparente e responsabilizável é dar a todos uma chance igual de receber o programa em primeiro, segundo ou terceiro lugar, e assim consecutivamente, implicando na implantação aleatória do programa.

Nos casos em que os recursos são limitados, isto é, nos quais nunca haverá recursos suficientes para alcançar a cobertura do programa (células A1 e A2, e B1 e B2), um excesso de demanda por esses recursos pode surgir muito rapidamente. Assim um sorteio para decidir quem fica no programa pode ser uma alternativa viável. Também nesse caso, todos terão a mesma chance de se beneficiar do programa. O sorteio é uma regra operacional justa, transparente e com atribuição de responsabilidades claras para alocar benefícios do programa.

Tabela 10.1 Relação entre as Regras Operacionais de um Programa e os Métodos de Avaliação de Impacto

CRONOGRAMA	RECURSOS →	**Excesso de demanda pelo programa (recursos limitados)**		**Sem excesso de demanda pelo programa (com todos os recursos)**	
	CRITÉRIO DE ELEGIBILIDADE DO PROGRAMA →	**Elegibilidade contínua ou classificação e ponto de corte (1)**	**Sem elegibilidade contínua ou classificação e ponto de corte (2)**	**Elegibilidade contínua ou classificação e ponto de corte (3)**	**Sem elegibilidade contínua ou classificação e ponto de corte (4)**
	Implementação gradual ao longo do tempo (A)	**CÉLULA A1** (3.1) Alocação aleatória (4) RDD	**CÉLULA A2** (3.1) Alocação aleatória (3.2) Promoção aleatória (5) DD com (6) Pareamento	**CÉLULA A3** (3.1) Alocação aleatória por fases (4) RDD	**CÉLULA A4** (3.1) Alocação aleatória por fases (3.2) Promoção aleatória para inclusão inicial (5) DD com (6) Pareamento
	Implementação imediata (B)	**CÉLULA B1** (3.1) Alocação aleatória (4) RDD	**CÉLULA B2** (3.1) Alocação aleatória (3.2) Promoção aleatória (5) DD com (6) Pareamento	**CÉLULA B3** (4) RDD	**CÉLULA B4** se menos que inclusão plena: (3.2) Promoção aleatória (5) DD com (6) Pareamento

Fonte: Autores.

Nota: O número entre parênteses refere-se ao capítulo do livro em que o método é discutido. RDD = do inglês *Regression Discontinuity Design*, Método de Regressão Descontinua; DD = diferenças-em-diferenças.

Outra classe de programas compreende aqueles que são aplicados gradualmente ao longo do tempo e para os quais os administradores podem classificar os potenciais beneficiários por necessidade — células A1 e A3. Se os critérios utilizados para priorizar os beneficiários forem quantitativos e disponíveis e tiverem um ponto de corte para a elegibilidade, o programa pode usar um desenho de regressão descontínua.

A outra grande categoria consiste em programas que têm a capacidade administrativa de ser implementados imediatamente — isto é, as células na

linha inferior do gráfico. Quando o programa tiver recursos limitados e não for capaz de classificar beneficiários (célula B2), pode-se usar a alocação aleatória com base no excesso de demanda. Se o programa tiver recursos suficientes para atingir escala mas sem critérios de focalização (célula B4), então a única solução será usar a promoção aleatória, supondo-se que nem todas as unidades elegíveis decidam participar do programa. Se o programa puder classificar beneficiários e for focalizado, pode-se usar novamente a regressão descontínua.

Encontrando o Grau Mínimo de Intervenção

As regras de funcionamento também determinam o grau mínimo de intervenção. O grau de intervenção é o grau em que o programa está sendo implementado. Por exemplo, se um programa de saúde é implementado em nível distrital, então todas as localidades do distrito ou receberiam o programa (como um grupo), ou não. Alguns programas podem ser executados de forma eficiente em nível individual, familiar ou institucional, enquanto outros precisam ser implementados em nível de distrito administrativo ou de comunidade. Implementar uma intervenção em um desses níveis mais elevados (por exemplo, uma província ou estado) pode ser problemático para a avaliação, por três razões principais:

1. O tamanho da amostra de avaliação e o custo da avaliação aumentam com o grau da intervenção.
2. À medida que o grau de intervenção aumenta, fica difícil encontrar um número suficiente de unidades para incluir na avaliação.
3. A validade interna da avaliação é mais suscetível de ser ameaçada por grandes unidades de intervenção.

Primeiro, as avaliações de intervenções implementadas em níveis mais altos, como de comunidade ou distrito administrativo, exigem amostras maiores e serão mais dispendiosas, em comparação a avaliações de intervenções em um nível inferior, como em nível individual ou familiar.[2] O nível de intervenção é importante porque define a unidade de alocaçãoaos grupos de tratamento e de comparação e isso tem implicações no tamanho da amostra de avaliação e no seu custo. Para intervenções implementadas em níveis mais altos, você precisará de uma amostra maior para conseguir detectar o verdadeiro impacto do programa. A lógica por trás disso será discutida no Capítulo 11, que revê os cálculos de poder estatístico e como estabelecer o tamanho da amostra necessária para uma avaliação.

Um ponto ligeiramente distinto é que o tamanho da amostra necessário para que a alocação aleatória tenha sucesso em equilibrar os grupos de tratamento e de comparação torna-se problemático em níveis elevados de agregação. Intuitivamente, se o nível de agregação for no nível estadual e o país tiver apenas seis estados, então é improvável que a aleatorização alcance o equilíbrio entre os grupos de tratamento e de comparação. Nesse caso, digamos que o desenho de avaliação aloque três estados para o grupo de tratamento e três para o grupo de comparação. É muito improvável que as condições no grupo de tratamento sejam semelhantes às do grupo de comparação, mesmo que o número de domicílios dentro de cada estado seja grande. A chave do equilíbrio dos grupos de tratamento e de comparação é o número de unidades atribuídas aos grupos de tratamento e de comparação e não o número de indivíduos ou domicílios na amostra.

O terceiro problema da utilização de unidades de intervenção de larga escala é que as alterações diferenciais ao longo do tempo são mais suscetíveis de comprometer a validade interna da alocação aleatória, mesmo que as características dos grupos estejam equilibradas na linha de base. Considere novamente o exemplo de usar os estados como o nível de intervenção de um programa de seguro saúde. A avaliação atribui aleatoriamente um grupo de estados ao grupo de tratamento e outro ao grupo de comparação. Suponha que você tenha sorte e que os dois grupos estejam equilibrados na linha de base, — isto é, os domicílios nos grupos de comparação e de tratamento têm, em média, o mesmo nível de despesas médicas. Depois da coleta dos dados da linha de base, alguns estados podem introduzir outras políticas de saúde por conta própria, tais como programas de vacinação ou de água e saneamento, que melhoram o estado de saúde da população e, dessa forma, diminuem a demanda por atendimento médico e os gastos com saúde. Se essas mudanças de políticas não forem niveladas entre os grupos de comparação e de tratamento, então o impacto do seguro saúde nos gastos com despesas de saúde será confundido com a mudança em outras políticas de saúde dos estados. De maneira similar, alguns estados podem experimentar crescimento econômico mais rápido do que outros. Despesas médicas muito provavelmente crescem mais rápido nos estados com crescimento de renda mais rápido. Novamente, se essas alterações diferenciais de crescimento econômico local não forem niveladas entre os grupos de comparação e de tratamento, então o impacto do seguro saúde sobre as despesas será confundido com a mudança na economia local. Em geral, é mais difícil controlar esses tipos de mudanças temporais em graus maiores de intervenção. Realizar alocação aleatória em pequenas unidades de implementação atenua essas ameaças à consistência interna.

Para evitar os problemas associados à implementação de uma intervenção em nível alto de unidade administrativa ou geográfica, os gestores do programa precisam encontrar o grau mínimo em que o programa possa ser implementado. Vários fatores determinam o grau mínimo viável de intervenção:

- Economias de escala e complexidade administrativa na execução do programa
- Capacidade administrativa de atribuir benefícios em nível individual ou domiciliar
- Preocupações sobre potenciais conflitos sociais
- Preocupações sobre a contaminação do grupo de comparação.

O grau mínimo de intervenção é, normalmente, baseado em economias de escala e na complexidade administrativa de se executar o programa. Por exemplo, um programa de seguro saúde pode exigir um escritório local para que os beneficiários enviem requerimentos e para pagar fornecedores. Os custos fixos do escritório precisam ser distribuídos entre um grande número de beneficiários, por isso pode ser ineficiente implantar o programa em nível individual e mais eficiente fazê-lo no nível da comunidade. No entanto, em situações com tipos de intervenções novas e não testadas, pode valer a pena absorver ineficiências de curto prazo e implantar o programa dentro de algumas unidades administrativas, de forma a melhor assegurar a credibilidade da avaliação e reduzir os custos de coleta de dados.

Alguns governos argumentam que programas administrados localmente, como programas de seguro saúde, não têm a capacidade administrativa para implantar programas em nível individual. Eles temem que seria um fardo criar sistemas para oferecer benefícios diferentes para diferentes beneficiários dentro de unidades administrativas locais e que o programa não seria capaz de garantir que a seleção de grupos de comparação e de tratamento seria implementada conforme concebida. Este último problema é uma séria ameaça à capacidade do governo de implementar o método de avaliação e, portanto, ao sucesso do estudo.

Às vezes, os governos preferem implementar programas em níveis mais agregados, como na comunidade, porque se preocupam com o potencial conflito social quando os membros do grupo de comparação observarem seus vizinhos no grupo de tratamento recebendo benefícios antes. Na realidade, pouca evidência foi apresentada para substanciar essas alegações. Um grande número de programas tem sido implementado com sucesso em nível individual ou domiciliar dentro das comunidades, sem gerar conflito social,

quando os benefícios são alocados de forma equitativa, transparente e com atribuição clara de responsabilidades.

Finalmente, quando um programa é implementado em um nível muito baixo, como no nível domiciliar ou individual, a contaminação do grupo de comparação pode comprometer a validade interna da avaliação. Por exemplo, digamos que você esteja avaliando o efeito do fornecimento de água de torneira na saúde das famílias. Se você instalar as torneiras de uma família, mas não as da a família do domicílio vizinho, a família do tratamento pode muito bem compartilhar o uso da torneira com seu vizinho de comparação; a família vizinha, então, não seria uma verdadeira comparação, uma vez que se beneficiaria de uma externalidade.

Portanto, na prática, os gestores do programa precisam encontrar o grau mínimo de intervenção que (1) permita uma amostra grande o suficiente para a avaliação; (2) atenue os riscos à validade interna, e; (3) se encaixe no contexto operacional. O Quadro 10.1 ilustra a escolha e as implicações do grau mínimo de intervenção no contexto dos programas de transferência de renda.

Quadro 10.1: Programas de Transferência de Renda e o Grau Mínimo de Intervenção

A maioria das transferências condicionadas de renda usam comunidades como o grau mínimo de intervenção, por razões administrativas e de desenho do programa, bem como por causa de preocupações envolvendo efeitos de transbordamentos e potencial conflito na comunidade, caso o tratamento fosse atribuído em um nível inferior.

Por exemplo, a avaliação do Progresa / Oportunidades, programa de transferência condicionada de renda do México, contou com a implantação do programa em nível comunitário em áreas rurais, para atribuir aleatoriamente comunidades aos grupos de tratamento e de comparação. A todos os domicílios elegíveis nas comunidades de tratamento foi oferecida a oportunidade de se inscrever no programa, na primavera de 1998, e a todos os domicílios elegíveis nas comunidades de comparação foi oferecida a mesma oportunidade 18 meses depois, no inverno de 1999. No entanto, os avaliadores encontraram uma forte correlação nos resultados entre domicílios dentro das comunidades. Por isso, para gerar poder estatístico suficiente para a avaliação, eles precisaram de mais domicílios na amostra do que teria sido necessário caso tivessem sido capazes de designar domicílios individuais aos grupos de tratamento e de comparação. Assim, a impossibilidade de implementar o programa em nível domiciliar suscitou a necessidade de usar de amostras maiores, aumentando o custo da avaliação. Restrições semelhantes se aplicam a uma grande parte de programas no setor de desenvolvimento humano.

Fontes: Behrman e Hoddinott 2001; Gertler 2004; Levy e Rodríguez 2005; Schultz 2004; Skoufias e McClafferty 2001.

A avaliação é ética?

Questões éticas são frequentemente levantadas sobre a condução de avaliações de impacto. Um ponto de partida para esse debate é considerar a ética de investir recursos públicos substanciais em programas cuja eficácia é desconhecida. Nesse contexto, a falta de avaliação pode, por si só, ser vista como antiética. As informações sobre a eficácia do programa que as avaliações de impacto geram podem levar ao investimento mais eficaz e ético dos recursos públicos.

Quando é tomada a decisão de desenhar uma avaliação de impacto, algumas questões éticas importantes devem ser consideradas. Referem-se às regras utilizadas para atribuir os benefícios do programa, bem como aos métodos pelos quais são realizados estudos envolvendo pessoas.

Conceito-chave:
Benefícios nunca devem ser negados ou atrasados somente para fins de avaliação. envolvendo pessoas.

O princípio mais básico da alocação dos benefícios do programa é que a entrega de benefícios nunca deve ser negada ou retardada com o propósito único da avaliação. Neste livro, argumentamos que a avaliação não deve ditar como os benefícios são alocados - ao contrário, as avaliações devem ser ajustadas às regras de seleção do programa. Nesse contexto, as preocupações éticas não decorrem da avaliação de impacto em si, mas diretamente das regras de seleção do programa.

A alocação aleatória de benefícios do programa frequentemente suscita questões éticas em torno da negação de benefícios do programa a beneficiários elegíveis. No entanto, a maioria dos programas opera com poucos recursos financeiros e administrativos, tornando impossível chegar a todos os beneficiários elegíveis ao mesmo tempo. Do ponto de vista ético, todos os indivíduos igualmente elegíveis para participar de qualquer tipo de programa social devem ter as mesmas chances de receber o programa. A alocação aleatória preenche este requisito ético. Em situações em que um programa será estendido gradualmente ao longo do tempo, a implantação pode ser baseada na seleção aleatória da ordem em que os beneficiários igualmente merecedores receberão o programa. Nesses casos, os beneficiários que entram no programa mais tarde podem ser usados como grupo de comparação para os primeiros beneficiários, gerando um desenho sólido de avaliação, bem como um método transparente e justo de alocar recursos escassos.

Em muitos países e instituições internacionais, foram criados conselhos de revisão ou comitês de ética para regulamentar as pesquisas envolvendo pessoas. Esses conselhos são incumbidos de avaliar, aprovar e monitorar estudos de pesquisa, com o principal objetivo de proteger os direitos e promover o bem-estar de todos os indivíduos pesquisados.

Embora as avaliações de impacto sejam, antes de mais nada, empreendimentos operacionais, também constituem estudos de pesquisa e, como tal, devem aderir às diretrizes de pesquisa envolvendo pessoas.

Nos Estados Unidos, o Escritório de Proteção à Pesquisa com Humanos do Departamento de Saúde e Serviços Humanos é responsável por coordenar e apoiar o trabalho dos conselhos de revisão institucional estabelecidos em todas as instituições de pesquisa e nas universidades. O Escritório de Proteção à Pesquisa com Humanos também publica uma compilação de mais de mil leis, regulamentos e diretrizes que regem a pesquisa com seres humanos em 96 países e fornece links para os códigos de ética e normas regulamentares atualmente utilizados pelas principais organizações internacionais e regionais.

Por exemplo, todas as pesquisas realizadas nos Estados Unidos ou financiadas por órgãos federais dos EUA, como o Instituto Nacional de Saúde e a Agência para o Desenvolvimento Internacional dos EUA, devem respeitar os princípios éticos e os requisitos regulamentares estabelecidos em lei federal.[3] Os princípios básicos da legislação dos EUA relativos à proteção dos seres humanos são baseados no histórico Relatório Belmont e incluem as garantias de que:

- a seleção de indivíduos seja equitativa;
- os riscos aos indivíduos sejam minimizados;
- os riscos aos indivíduos sejam razoáveis em relação aos benefícios esperados;
- seja obtido consentimento informado de cada potencial participante ou de seu representante legal;
- as disposições adequadas estejam em vigor para proteger a privacidade dos indivíduos e manter a confidencialidade; e
- que garantias adicionais sejam incluídas para indivíduos mais vulneráveis, como crianças, prisioneiros e os economicamente desfavorecidos.

Embora a lista decorra da experiência histórica com testes médicos, os princípios básicos de proteção aos direitos e promoção do bem-estar de todos os indivíduos são, atualmente, aplicáveis à pesquisa social. No contexto da avaliação de programas sociais, os primeiros três pontos se relacionam à ética de alocação de benefícios. Os últimos três pontos se referem aos protocolos com base nos quais os indivíduos são estudados para fins de avaliação.[4]

Ao desenhar, gerenciar ou contratar uma avaliação de impacto, você deve certificar-se de que todos os estágios tenham respaldo em alguma lei ou processo de revisão existente que regule pesquisas com seres humanos, seja do país onde a avaliação é implementada, seja do país onde a agência de financiamento está localizada.

Como Compor uma Equipe de Avaliação?

Conceito-chave:
Uma avaliação é uma parceria entre os formuladores de políticas e os avaliadores.

Uma avaliação é uma parceria entre os formuladores de políticas e os avaliadores, com cada grupo dependendo do outro para o seu sucesso. Os formuladores de políticas são responsáveis por orientar o trabalho e garantir a relevância da avaliação — formulando as questões de avaliação, determinando se uma avaliação de impacto é necessária, supervisionando a avaliação, assegurando recursos adequados para o trabalho e aplicando os resultados. Os avaliadores são responsáveis pelos aspectos técnicos — a metodologia de avaliação, o desenho amostral, a coleta de dados e a análise.

Uma avaliação é um equilíbrio entre a especialidade técnica e independência trazidas por um grupo externo de avaliadores e a relevância da política, orientação estratégica e coordenação operacional aportadas pelos formuladores de políticas. Nessa parceria, um elemento-chave é determinar que grau de separação institucional estabelecer entre os realizadores da avaliação e os usuários da avaliação. Muito pode ser obtido com a objetividade decorrente de ter uma avaliação realizada com independência da instituição responsável pelo projeto sendo avaliado. No entanto, as avaliações podem, muitas vezes, ter objetivos múltiplos, incluindo a construção de capacidade de avaliação nas agências governamentais e a sensibilização de operadores do programa às realidades de seus projetos, uma vez executados no campo.

Para uma avaliação de impacto ser bem sucedida, os avaliadores e os formuladores de políticas devem trabalhar juntos. Se, por um lado, as avaliações de impacto devem ser realizadas por um grupo externo para manter a objetividade e credibilidade, por outro, o processo não pode ser dissociado das regras operacionais. Em particular, quando se trata da apreciação das regras de implementação do programa, para determinar o desenho mais adequado da avaliação e garantir que a implementação e a avaliação do programa estejam bem coordenadas, de modo que não comprometam uma a outra. Além disso, os resultados têm menor probabilidade de serem diretamente relevantes à política ou de terem impacto na política se não houver o envolvimento dos formuladores de políticas desde o início.

A Composição de uma Equipe de Avaliação

Os formuladores de políticas podem solicitar uma avaliação utilizando várias formas de contratação. Primeiro, a unidade do governo que contrata a avaliação pode decidir contratá-la por completo, de uma só vez. Nesse caso, ela será responsável por estabelecer pelo menos um primeiro esboço do plano de avaliação, incluindo os objetivos-chave, as questões de política, a metodologia prevista, os dados a ser coletados e os tetos orçamentários. Este plano prevê os termos básicos de referência para lançar um edital de propostas técnicas e financeiras de avaliadores externos. Os termos também podem especificar uma composição mínima de equipe que os avaliadores externos terão de cumprir. A preparação de propostas técnicas dá aos avaliadores externos a chance de sugerir melhorias ao plano de avaliação que o governo produziu. Uma vez que a avaliação seja contratada, a agência externa que foi contratada gerencia ativamente a avaliação e designa um gestor da avaliação. Nesse modelo, a equipe do governo proporciona, principalmente, a supervisão.

Sob um segundo tipo de formato contratual, a unidade de governo que contrata a avaliação pode decidir geri-la diretamente. Isso requer o desenvolvimento de um plano de avaliação de impacto e a subsequente contratação de seus subcomponentes. Neste formato, o gerente da avaliação permanece na unidade de governo que contrata a avaliação.

Independentemente do formato contratual, uma responsabilidade básica do gerente de avaliação é constituir a equipe de avaliação, tendo em mente os interesses dos clientes e os passos necessários para realizá-la. Embora cada avaliação seja diferente, a equipe técnica de qualquer esforço de avaliação de impacto que se apoie na coleta de seus próprios dados, qualitativos ou quantitativos, quase sempre precisará de alguns membros específicos, incluindo:

- *Um gestor de avaliação.* Essa pessoa é responsável por estabelecer os objetivos-chave, as questões de política, os indicadores e as necessidades de informação da avaliação (frequentemente em estreita colaboração com os formuladores de políticas e usando uma teoria de mudança, como a cadeia de resultados); selecionar a metodologia de avaliação; identificar a equipe de avaliação; e elaborar termos de referência para as partes da avaliação a serem contratadas ou subcontratadas. É importante designar um gestor da avaliação que seja capaz de trabalhar de forma eficaz com os produtores de dados, bem como com os analistas e formuladores de políticas que usam os dados e os resultados da avaliação. Se a pessoa não residir no local, recomenda-se que um gestor local seja designado para coordenar o esforço de avaliação em conjunto com o gestor internacional.

- *Um especialista em amostragem*. Essa pessoa é alguém que possa orientar o trabalho de cálculos de poder estatístico e de amostragem. Para as avaliações quantitativas de impacto, o especialista em amostragem deve ser capaz de realizar cálculos de poder estatístico para determinar o tamanho das amostras apropriadas aos indicadores estabelecidos, selecionar a amostra, rever os resultados da amostra real comparados à amostra prevista, e aconselhar tecnicamente a análise, por exemplo, sobre como incorporar os pesos amostrais, se necessário. O especialista de amostragem também deverá ter a tarefa de selecionar os locais e os grupos para o teste piloto. Principalmente se o especialista de amostragem for um consultor internacional, ele ou ela frequentemente terá de formar dupla com um coordenador local de informações, responsável pela coleta de dados a partir dos quais a amostra será extraída.
- *Uma pessoa ou equipe responsável pela concepção dos instrumentos de coleta de dados e acompanhamento de manuais e livros de códigos.* Esta pessoa trabalha com o gerente de avaliação para assegurar que os instrumentos de coleta de dados, de fato, produzirão os dados necessários para a análise; esta pessoa ou equipe também participa dos testes piloto dos questionários.
- *Uma equipe de trabalho de campo*. A equipe inclui um gerente de trabalho de campo que possa supervisionar todo o esforço de coleta de dados, desde o planejamento das rotas para coleta de dados até a formação e programação das equipes de campo, que são geralmente compostas por supervisores e entrevistadores.
- *Gerentes de dados e técnicos em processamento de dados.* Eles projetam os programas de entrada de dados, entram os dados, verificam sua validade, fornecem a documentação necessária dos dados e produzem os resultados básicos para verificação pelos analistas de dados.
- *Analistas de dados e de políticas.* Os analistas trabalham com os dados produzidos e com o gestor da avaliação para realizar a análise necessária e produzir os relatórios de avaliação.

Parceiros de Avaliação

Uma das primeiras determinações que os formuladores de políticas, juntamente com o gestor da avaliação, devem fazer é se a avaliação — ou partes dela —, pode ser implementada localmente e que tipo de supervisão e assistência externa será necessária. A capacidade de avaliação varia muito de país para país. Contratos internacionais que permitem que empresas de um

país realizem avaliações em outro país estão se tornando mais comuns. Também está se tornando cada vez mais comum governos e instituições multilaterais implementarem avaliações localmente, ao mesmo tempo em que se oferece uma boa dose de supervisão internacional. Cabe ao gestor da avaliação avaliar criticamente a capacidade local e determinar quem será responsável por quais aspectos do esforço de avaliação.

Outra questão é a de se trabalhar com uma empresa privada ou órgão público. Empresas privadas ou instituições de pesquisa podem ser mais confiáveis para fornecer resultados em tempo hábil, mas se perde a capacitação do setor público. Além disso, frequentemente as empresas privadas são, compreensivelmente, menos receptivas a incorporar elementos na avaliação que encareçam o esforço. Instituições de pesquisa e universidades também podem funcionar como avaliadores. A especialidade técnica e reputação de instituições sólidas de pesquisa ou universidades podem garantir que os resultados da avaliação sejam amplamente aceitos pelas partes interessadas. No entanto, a essas instituições algumas vezes falta a experiência operacional ou a capacidade de executar alguns aspectos da avaliação - como a coleta de dados - de modo que esses aspectos podem ter de ser subcontratados a outro parceiro. Independentemente da combinação de parceiros, um exame cuidadoso das atividades passadas dos potenciais colaboradores é essencial para que seja feita uma escolha informada.

Especialmente quando ao trabalhar com um órgão público, um avaliador zeloso deve ter consciência da capacidade da equipe de avaliação, à luz de outras atividades que a unidade esteja realizando. Isto é particularmente relevante quando se trabalha com órgãos do setor público com múltiplas responsabilidades e recursos humanos limitados. A consciência da carga de trabalho da unidade é importante para avaliar não só como ela afetará a qualidade da avaliação sendo realizada, mas também o custo de oportunidade da avaliação com respeito a outras tarefas pelas quais a unidade é responsável. Num exemplo, foi planejada uma avaliação de impacto de uma reforma educacional que exigia esforços do pessoal da equipe de avaliação nacional responsável pelos testes nacionais de desempenho semestral. A equipe foi selecionada como parte do esforço de avaliação porque era a mais qualificada profissionalmente para assumir a responsabilidade pela avaliação e porque se buscavam complementaridades entre a avaliação e o exame nacional. No entanto, quando a reforma — e sua avaliação correspondente — foi adiada, o atraso descarrilou todo o esforço de pesquisa; os testes de desempenho para a avaliação nacional não foram aplicados conforme o cronograma e o país perdeu uma oportunidade de monitorar o progresso educacional. Tais situações podem ser evitadas

através da coordenação com os gerentes na unidade responsável pela avaliação, para garantir que seja atingido um equilíbrio na programação de diversas atividades, bem como na distribuição de pessoal e recursos entre essas atividades.

Como Programar a Avaliação?

Discutimos na Parte 1 as vantagens de avaliações prospectivas, concebidas durante a preparação do programa. O planejamento antecipado permite uma escolha mais ampla na geração de grupos de comparação, facilita a coleta de dados de linha de base e ajuda os interessados a chegar a um consenso sobre os objetivos do programa e as questões de interesse.

Embora seja importante planejar as avaliações no início da fase de concepção do projeto, sua execução deve ser programada para avaliar o programa quando este já esteja amadurecido. Os projetos-piloto ou reformas incipientes são frequentemente propensos à revisão, tanto de seu conteúdo quanto em relação ao *como*, *quando*, *onde* e *por quem* serão implementadas. Os implementadores do programa podem precisar de tempo para aprender e aplicar novas regras operacionais consistentemente. Visto que as avaliações requerem regras claras de funcionamento do programa para gerar contrafactuais adequados, é importante realizar as avaliações dos programas depois que estes já estejam bem consolidados.

Os dados de linha de base devem sempre ser coletados, mas outra questão-chave referente ao cronograma é quanto tempo será necessário antes de que os resultados possam ser medidos. O equilíbrio é bastante vinculado ao contexto: "se a avaliação for cedo demais, existe o risco de se encontrar apenas um impacto parcial ou impacto nenhum; se feita demasiadamente tarde, haverá o risco de que o programa possa perder doações e apoio público, bem como o risco de expansão de um programa mal concebido" (King e Behrman 2009, p. 56). Os seguintes fatores precisam ser ponderados para determinar o tempo de coleta dos dados de seguimento:[5]

- O ciclo do programa, incluindo a duração, tempo de execução e potenciais atrasos
- O tempo previsto para que o programa influencie os resultados, assim como a natureza dos resultados de interesse
- Ciclos de formulação de políticas.

Primeiro, a avaliação de impacto precisa ser adaptada ao ciclo de execução do programa. A avaliação não pode reger o programa sendo avaliado.

Por sua própria natureza, as avaliações estão sujeitas aos prazos do programa; devem estar alinhadas à duração prevista do programa. Também devem ser adaptadas a possíveis defasagens de execução, quando os programas demoram a entregar benefícios ou atrasam devido a fatores externos.[6] Em geral, embora o cronograma de avaliação deva ser incorporado ao projeto desde o início, os avaliadores deverão estar preparados para ser flexíveis e realizar modificações à medida que o projeto é executado. Além disso, medidas devem ser tomadas para realizar o acompanhamento das intervenções, utilizando um sistema de monitoramento robusto, de maneira que esforço de avaliação seja informado pelo ritmo real da intervenção.

A coleta de dados de seguimento deve levar em conta quanto tempo é necessário após a implementação do programa para que os resultados se tornem aparentes. A cadeia de resultados do programa auxilia na identificação dos indicadores de resultados e do momento adequado de mensuração. Alguns programas (como programas de apoio à renda) têm como objetivo proporcionar benefícios de curto prazo, enquanto outros (como programas de educação básica) objetivam ganhos de mais longo prazo. Além do mais, certos resultados, por sua natureza, demoram mais a aparecer (como mudanças na expectativa de vida ou na fertilidade, resultantes de uma reforma no sistema da saúde) do que outros (como a renda após um programa de capacitação).

Na avaliação do Fundo de Investimentos Sociais da Bolívia, por exemplo, que se apoiou em dados de linha de base coletados em 1993, os dados de seguimento não foram coletados até 1998, por causa do tempo necessário para que as intervenções fossem realizadas (água e saneamento, postos de saúde e escolas) e para que surgissem os impactos na saúde e educação da população beneficiária (Newman et al. 2002). Foi necessário um período de tempo similar para a avaliação de um projeto de educação primária no Paquistão, em que se utilizou um método experimental com pesquisas de linha de base e de seguimento, para avaliar o impacto de escolas comunitárias sobre os resultados dos alunos, incluindo o desenvolvimento acadêmico (King, Orazem e Paterno 2008).

O momento da coleta de dados de seguimento depende, portanto, do programa em estudo, bem como dos indicadores de resultados de interesse. Algumas avaliações terão dados de seguimento coletados enquanto o programa ainda estiver sendo executado, para medir mudanças de curto prazo e para manter contato com a amostra de avaliação, visando à redução do atrição da amostra ao longo do tempo. Para programas sem operações contínuas, ciclos adicionais de dados de seguimento, coletados bem depois da conclusão do programa, podem ajudar a mensurar os impactos

de longo prazo. Dados de seguimento podem ser coletados mais de uma vez, de forma que os impactos de curto e médio prazo possam ser considerados e contrastados.

O *seguimento* de dados coletados durante a execução do programa pode não capturar todo o impacto do programa, se os indicadores forem medidos cedo demais. De fato, "os programas não necessariamente alcançam um grau estável de eficácia logo depois que a implementação começa. O aprendizado pelos provedores e beneficiários pode levar tempo" (King e Behrman 2009, p. 65). Ainda assim, é muito útil documentar impactos de curto prazo. Como já foi afirmado, alguns programas têm apenas objetivos de curto prazo (como o apoio à renda). A evidência de como um programa deste tipo funciona no curto prazo também pode fornecer informações sobre resultados previstos de mais longo prazo. Por exemplo, é importante medir frequentemente os indicadores de curto prazo que são bons preditores de indicadores de longo prazo (tais como os partos assistidos como indicador de curto prazo da mortalidade infantil). Dados de *seguimento* coletados enquanto o programa ainda está sendo implementado também são úteis para a produção de resultados de avaliação do impacto inicial, que podem fortalecer o diálogo entre avaliadores e formuladores de políticas.

Pesquisas de *seguimento,* que medem resultados de longo prazo depois da implementação do programa, muitas vezes produzem a evidência mais convincente sobre a efetividade do programa. Por exemplo, os resultados positivos das avaliações de impacto de longo prazo de programas para a primeira infância nos Estados Unidos (Currie e Thomas 1995, 2000; Currie 2001) e na Jamaica (Grantham-McGregor et al. 1994) têm sido influentes na defesa de investimentos para intervenções na primeira infância.

Impactos de longo prazo algumas vezes constituem objetivos de programa explícitos, mas também podem refletir efeitos não intencionais, indiretos, como aqueles relacionados a mudanças de comportamento. A identificação de impactos de longo prazo, no entanto, pode criar dificuldades. Impactos podem simplesmente desaparecer no longo prazo. Mesmo um método robusto de avaliação de impacto pode não resistir ao teste do tempo. Por exemplo, as unidades do grupo de controle podem começar a se beneficiar das externalidades dos beneficiários do programa.

Apesar dos dados de seguimento de curto prazo e de longo prazo serem complementares, a programação de uma avaliação também deve levar em conta os momentos quando certas informações são necessárias para a tomada de decisão e devem sincronizar atividades de coleta de dados e avaliação com momentos-chave do processo de tomada de decisão. A produção de resultados deve ser programada para informar orçamentos, programar expansões ou outras decisões de políticas.

Como Orçar a Avaliação?

A orçamentação constitui uma das últimas etapas de operacionalização do desenho da avaliação. Nesta seção, examinaremos alguns dados referentes a custos de avaliações de impacto, discutiremos como orçar uma avaliação e sugeriremos algumas opções de financiamento.

Revisão de Dados de Custos

As Tabelas 10.2 e 10.3 contêm dados de custo de avaliações de impacto de uma série de projetos apoiados pelo Banco Mundial. A amostra na Tabela 10.2 é decorrente de uma ampla revisão de programas apoiados pela unidade de Proteção Social e Trabalho. A amostra na Tabela 10.3 foi selecionada com base na disponibilidade de estatísticas orçamentárias atuais acerca do conjunto de avaliações de impacto financiado pelo Fundo Estratégico de Avaliação de Impacto (SIEF, do inglês *Strategic Impact Evaluation Fund*).

Tabela 10.2 Custos de Avaliações de Impacto de uma Seleção de Projetos Apoiados pelo Banco Mundial

Avaliação de impacto	País	Custo total da AI ($)	Custo total do programa ($)	% AI dos custos totais do programa
Emprego e Desenvolvimento de Competências do Migrante	China	220.000	50.000.000	0,4
Projeto de Redes de Proteção Social	Colômbia	130.000	86.400.000	0,2
Programa de Investimento em Setores Sociais	República Dominicana	600.000	19.400.000	3,1
Proteção Social	Jamaica	800.000	40.000.000	2,0
Assistência Técnica ao Projeto Redes de Proteção Social	Paquistão	2.000.000	60.000.000	3,3
Projeto de Proteção Social	Panamá	1.000.000	24.000.000	4,2
1º Projeto de Padrão de Vida Comunitária	Ruanda	1.000.000	11.000.000	9,1
Fundo Social de Desenvolvimento 3	Iêmen, Rep.	2.000.000	15.000.000	13,3
Média		**968.750**	**38.225.000**	**4,5**

Fonte: Cálculos dos autores, a partir de uma amostra de programas do Banco Mundial no Setor de Proteção Social

Nota: AI = avaliação de impacto.

Tabela 10.3 Custos Desagregados de uma Seleção de Projetos Apoiados pelo Banco Mundial

			Desagregação dos Custos do AI				
Avaliação de impacto SIEF	**País**	**Custo total**	**Viagens**	**Pessoal do Banco Mundial**	**Consultores (nacionais e intern.)**	**Coleta de dados (incluindo pessoal de campo)**	**Outros (divulgacao e oficinas**
Crédito de Apoio à Redução da Pobreza e Saúde Materna	Benin	1.690.000	270.000	200.000	320.000	840.000	60.000
Reforma de Pagamentos por Desempenho aos Professores	Brasil	513.000	78.000	55.000	105.000	240.000	35.000
Programa *Nadie es Perfecto* para Desenvolver Competências Parentais	Chile	313.000	11.500	—	260.000	35.500	6.000
Pagamento por Desempenho no Setor da Saúde na China: Avaliação de Saúde XI	China	308.900	60.000	35.000	61.000	152.900	—
Programa Nacional de Garantia de Emprego Rural	Índia	390.000	41.500	50.000	13.500	270.000	15.000
Saúde e Nutrição Escolar: o Papel do Controle da Malária na Melhoria da Educação	Quênia	652.087	69.550	60.000	103.180	354.000	65.357
Campanha de Prevenção do HIV para a Juventude: Abstinência, Fidelidade e Sexo Seguro	Lesoto	630.300	74.300	9.600	440.000	98.400	8.000
PTCR, Escolarização e Risco de HIV	Malawi	1.842.841	83.077	144.000	256.344	1.359.420	—

Programa *Contigo Vamos por Mas Oportunidades*, no Estado de Guanajuato	México	132.199	2.660	50.409	—	80.640	1.150
Piloto PTCR Aleatorizado no Ensino Primário Rural	Marrocos	674,367	39,907	66,000	142,460	426,000	—
Aprender e Crescer na Sombra do HIV / AIDS: Programa ECD Aleatorizado	Moçambique	838,650	86,400	31,000	62,500	638,750	20,000
Formação de Distribuidores Comunitários na Prevenção e Tratamento da Malária	Nigéria	1,024,040	64,000	35,000	106,900	817,740	—
Saúde Escolar e Nutrição: o Papel do Controle da Malária na Melhoria da Educação	Senegal	644,047	61,800	60,000	102,890	354,000	65,357
PTCR para Prevenir o HIV e Outras Infecções Sexualmente Transmissíveis	Tanzânia	771,610	60,000	62,000	100,000	518,611	30,999
Média		**744,646**	**71,621**	**66,031**	**115,975**	**482,290**	**30,686**

Fonte: Cálculos dos autores, a partir de uma amostra de avaliações de impacto financiados pelo Fundo Estratégico de Avaliação de Impacto.

Nota: PTCR = programa de transferência condicionada de renda; ECD = desenvolvimento da primeira infância (do inglês *early childhood development*) = não disponível.

Apesar das duas amostras não serem necessariamente representativas de todas as avaliações realizadas pelo Banco Mundial, como os dados de custos ainda não foram documentados consistentemente, tais avaliações fornecem pontos de referência úteis sobre os custos associados à realização de avaliações de impacto rigorosas.

Os custos diretos das atividades de avaliação variam entre $130 mil e $2 milhões, com um custo médio de $968.750. Embora esses custos variem muito e possam parecer altos em termos absolutos, em termos relativos eles representam entre 0,2% e 13,3% dos custos totais do programa,[7] com uma média de 4,5%. Com base nessa amostra, avaliações de impacto constituem apenas uma pequena porcentagem do orçamento total do programa. Além disso, o custo da realização de uma avaliação de impacto deve ser comparado aos custos de oportunidade da não realização de uma avaliação rigorosa e, portanto, de execução de um programa potencialmente ineficaz. As avaliações permitem aos pesquisadores e formuladores de políticas identificarem quais programas ou aspectos dos programa funcionam, quais não, além de quais estratégias podem ser as mais eficazes e eficientes para se atingir as metas do programa. Nesse sentido, os recursos necessários para implementar uma avaliação de impacto constituem um investimento relativamente pequeno, mas significativo.

A Tabela 10.3 desagrega os custos da amostra de avaliações de impacto apoiados pelo SIEF. Os custos totais de uma avaliação incluem o tempo da equipe do Banco Mundial, consultores nacionais e internacionais, viagens, coleta de dados e atividades de divulgação.[8] Nessas, assim como em quase todas as avaliações para as quais não podem ser utilizados dados existentes, o maior custo é a coleta de dados novos, representando, em média, mais de 60% do custo.

É importante ter em mente que esses números refletem avaliações de diferentes tipos e amplitudes. O custo relativo da avaliação de um programa piloto é, geralmente, maior do que o custo relativo da avaliação de um programa de âmbito nacional ou universal. Além disso, algumas avaliações exigem apenas uma pesquisa de seguimento ou podem valer-se de fontes de dados existentes, enquanto outras podem necessitar de várias rodadas de coleta de dados. O Manual de Estudos de Medição de Padrões de Vida (do inglês, *The Living Standards Measurement Study Manual)* (Grosh e Glewwe 2000) fornece estimativas do custo da coleta de dados através de pesquisas domiciliares, baseado em experiências em países de todo o mundo. No entanto, o manual também ressalta que os custos dependem, em grande parte, das habilidades da equipe local, dos recursos disponíveis e da duração do período em campo. Para saber mais sobre

como orçar uma pesquisa em um contexto específico, recomenda-se que os avaliadores entrem em contato, inicialmente, com o instituto nacional de estatística.

Orçamento para uma Avaliação de Impacto

Claramente, são necessários muitos recursos para implementar uma avaliação de impacto rigorosa. Itens do orçamento incluem custos de pessoal para, pelo menos, um investigador / pesquisador principal, um assistente de pesquisa, um coordenador de campo, um especialista em amostragem, entrevistadores e pessoal do projeto, que podem oferecer apoio durante a avaliação. Esses recursos humanos podem constituir-se de pesquisadores e técnicos de organizações internacionais, consultores internacionais ou locais e pessoal do programa do país cliente. As despesas de viagem e de subsistência (hotéis e diárias) também devem ser orçadas. Recursos para a divulgação, muitas vezes na forma de oficinas, relatórios e trabalhos acadêmicos, também devem ser considerados no planejamento da avaliação.

Como dissemos, os maiores custos de uma avaliação são, geralmente, os referentes à coleta de dados (incluindo a criação e o teste piloto da pesquisa), aos materiais e equipamentos de coleta de dados, treinamento para os entrevistadores, pagamento de diárias, veículos e combustível e operações de entrada de dados. Calcular os custos de todos esses insumos requer algumas suposições sobre, por exemplo, quanto tempo será necessário para completar o questionário e o tempo de viagem entre os locais de pesquisa. Uma planilha de trabalho é apresentada na Tabela 10.4, para ajudar com a estimativa dos custos na etapa de coleta de dados.

Os custos de uma avaliação de impacto podem ser distribuídos ao longo de vários anos fiscais. Um exemplo de orçamento na Tabela 10.5 mostra como os gastos em cada fase de uma avaliação podem ser distribuídos pelos anos fiscais, para efeito da contabilidade e dos relatórios. Mais uma vez, as demandas orçamentárias provavelmente serão maiores durante os anos em que os dados forem coletados.

Fundos para Avaliações

Os recursos financeiros para a avaliação podem vir de várias fontes, incluindo empréstimos de projeto, orçamentos diretos de programas, bolsas de pesquisa ou patrocínio de doadores. Muitas vezes, as equipes de avaliação buscam uma combinação de fontes para gerar os fundos necessários. Embora os fundos para avaliações costumassem sair, principalmente, de orçamentos para pesquisas, uma ênfase crescente na formulação de políticas baseada em

Tabela 10.4 Planilha para Astimativa de Custo de uma Avaliação de Impacto

Tarefas e Recursos	Número	Preço / Unidade	Nº de Unidades	Total
Pessoal				
Equipe de avaliação de programas (gerente de avaliação, etc.)				
Consultores nacionais e/ou internacionais (pesquisador / investigador principal)				
Assistente de pesquisa				
Especialista em estatística				
Coordenador de campo				
Viagens				
Passagens aéreas internacionais e locais				
Transporte terrestre local				
Subsistência (hotel e diárias)				
Coleta de Dados[a]				
Desenho do instrumento				
Piloto				
Treinamento				
Viagens e diárias				
Material, equipamento de pesquisa				
Questionários impressos				
Pessoal de campo				
Entrevistadores				
Supervisores				
Transporte (veículos e combustível)				
Motoristas				
Entrada de dados e limpeza				
Análise de dados e disseminação				
Oficinas / workshops				
Artigos, relatórios				
Outros				
Escritório				
Comunicações				
Software				

Fonte: Autores.

a. Os cálculos de coleta de dados devem refletir pressupostos como o número de rodadas de coleta de dados necessárias, quanto tempo a coleta de dados levará, o número de comunidades na amostra, o número de domicílios por comunidade, o tamanho do questionário, o tempo de viagem e assim por diante.

Tabela 10.5 Exemplo de um Orçamento de Avaliação de Impacto

	Estágio de Desenho				Estágio do Dados de Linha de Base			
	Unidade	**Custo por unidade ($)**	**Nº de unidades**	**Custo total ($)**	**Unidade**	**Custo por unidade ($)**	**Nº de unidades**	**Custo total ($)**
A. Salários de funcionários	Semanas	7.500	2	15.000	Semanas	7.500	2	15.000
B. Honorários de consultores				10.250				27.940
Consultor Internacional (1)	Dias	45	15	6.750	Dias	45	0	0
Consultor Internacional (2)	Dias	35	10	3.500	Dias	35	10	3.500
Assistente de pesquisa / Coordenador de campo	Dias	18	0	0	Dias	18	130	24.440
C. Viagens e subsistência				14.100				15.450
Equipe: passagem aérea internacional	Viagens	3.350	1	3.350	Viagens	3.350	1	3.350
Equipe: hotel & diárias	Dias	15	5	750	Dias	15	5	750
Passagens aéreas internacionais: consultores internacionais	Viagens	3.500	2	7.000	Viagens	3.500	2	7.000
Hotel & diárias: consultores internacionais	Dias	150	20	3.000	Dias	150	20	3.000
Passagens aéreas internacionais: coordenador de campo	Viagens		0	0	Viagens	1.350	1	1.350
Hotel & diárias: consultores internacionais	Dias		0	0	Dias	150	0	0
D. Coleta de dados								126.000
Dados tipo 1: consentimento					Escola	12	100	12.000
Dados tipo 2: resultados da educação					Criança	1	3.000	42.000
Dados tipo 3: resultados de saúde					Criança	2	3.000	7.200
V. Outro(s)								
Oficina(s)								
Divulgação/ relatórios								
Outros 1 (custos fixos de coordenação em nível de conglomerado)								
Total de custos por estágio	Estágio de desenho:			39.350	Estágio de linha de base:			184.390

(continuação)

Tabela 10.5 *(continuação)*

	Seguimento de Dados Estágio I				Seguimento de Dados Estágio II			
	Unidade	**Custo por unidade ($)**	**Nº de unidades**	**Custo total ($)**	**Unidade**	**Custo por unidade ($)**	**Nº de unidades**	**Custo total ($)**
A. Salários de funcionários	Semanas	7.500	2	15.000	Semanas	7.500	2	15.000
B. Honorários de consultores				32.550				32.440
Consultor Internacional (1)	Dias	450	15	6.750	Dias	45	10	4.500
Consultor Internacional (2)	Dias	350	20	7.000	Dias	35	10	3.500
Assistente de pesquisa / Coordenador de campo	Dias	188	100	18.800	Dias	18	130	24.440
C. Viagens e subsistência				20.000				20.000
Equipe: passagem aérea internacional	Viagens	3.350	2	6.700	Viagens	3.350	2	6.700
Equipe: hotel & diárias	Dias	150	10	1.500	Dias	15	10	1.500
Passagens aéreas internacionais: consultores internacionais	Viagens	3.500	2	7.000	Viagens	3.500	2	7.000
Hotel & diárias: consultores internacionais	Dias	150	20	3.000	Dias	150	20	3.000
Passagem aérea internacional: coordenador de campo	Viagens	1.350	1	1.350	Viagens	1.350	1	1.350
Hotel & diárias: consultores internacionais	Dias	150	3	450	Dias	15	3	450
D. Coleta de dados				114.000				114.000
Dados tipo 1: consentimento								
Dados tipo 2: resultados da educação	Criança	14	3.000	42.000	Criança	1	3.000	42.000
Dados tipo 3: resultados de saúde	Criança	24	3.000	72.000	Criança	2	3.000	72.000
V. Outros								**65.357**
Oficina(s)						20.000	2	40.000
Divulgação / relatórios						5.000	3	15.000
Outros 1 (custos fixos de coordenação em nível de conglomerado)						5.179	2	10.357
Total de custos por estágio	Seguimento estágio I			181.550	Seguimento estágio II			246.797
							Total de custos da avaliação:	652.087

Fonte: Autores.

evidências tem aumentado os financiamentos vindos de outras fontes. Nos casos em que a avaliação puder preencher uma grande lacuna de conhecimento e que seja de interesse mais amplo para a comunidade de desenvolvimento - e onde puder ser aplicada uma avaliação com credibilidade e robusta - os formuladores de políticas devem ser motivados a buscar financiamento externo, dado o bem público representado pelos resultados da avaliação. Fontes de financiamento incluem o governo, bancos de desenvolvimento, organizações multilaterais, organismos das Nações Unidas, fundações, filantropos e organizações de pesquisa e avaliação, como a Iniciativa Internacional para a Avaliação de Impacto.

Notas

1. Responsabilização é a tradução que foi escolhida pelos tradutores deste livro para o termo em Inglês *accountability*, que captura tanto a idéia de responsabilização quanto de o de prestação de contas. *Accountability* sintetiza a concepção de que as pessoas que desempenham funções públicas deve regularmente explicar o que faz, como faz, por que faz, quanto gasta e o que planeja fazer.
2. A discussão nesta seção se aplica mais diretamente a um desenho de alocação aleatória, mas os mesmos princípios se mantêm inalterados para avaliações baseadas em outras metodologias.
3. Veja Kimmel 1988; NIH 2006; USAID 2008; U.S. Departamento de Saúde e Serviços Humanos 2010 e Arquivo Nacional dos EUA de 2009.
4. Os potenciais riscos na coleta de dados para a avaliação de programas sociais incluem: não obter o consentimento informado dos pesquisados; testar o desenvolvimento cognitivo de crianças na presença de seus pais, o que pode levar a hipóteses sobre as capacidades futuras das crianças; pedir para falar a sós com mulheres ou entrevistar mulheres sobre assuntos sensíveis na presença de familiares do sexo masculino; não compreender o momento ou o custo de oportunidade de entrevistar as pessoas e proporcionar uma compensação ou um gesto de agradecimento, quando apropriado.
5. Vide King e Behrman (2009) para uma discussão detalhada das questões de tempo em relação à avaliação de programas sociais.
6. "Existem várias razões pelas quais a implementação não é nem imediata nem perfeita, por que o tempo de exposição a um tratamento difere não só em áreas do programa, mas também entre os beneficiários finais e por que diferentes tempos de exposição podem levar a diferentes estimativas de impacto do programa" (King e Behrman 2009, 56).
7. Nesse caso, o custo é calculado como uma percentagem da parte do custo do projeto financiado pelo Banco Mundial.
8. Esse montante não inclui os custos de pessoal do projeto local que, com frequência, participaram proeminentemente do desenho e da supervisão da avaliação, uma vez que dados precisos sobre esses custos não são registrados regularmente.

Referências

Behrman, J. & Hoddinott, J. (2001). An Evaluation of the Impact of PROGRESA on Pre-school Child Height. *FCND Briefs 104*. Washington, DC: Instituto Internacional de Pesquisa sobre as Políticas Alimentares.

Currie, J. (2001). Early Childhood Education Programs. *Journal of Economic Perspectives* 15 (2): 213–38.

Currie, J. & Thomas, D. (1995). Does Head Start Make a Difference? *American Economic Review* 85 (3): 341–64.

——. (2000). School Quality and the Longer-Term Effects of Head Start. *Journal of Economic Resources* 35 (4): 755–74.

Gertler, P. (2004). Do Conditional Cash Transfers Improve Child Health? Evidence from PROGRESA's Control Randomized Experiment. *American Economic Review* 94 (2): 336–41.

Grantham-McGregor, S., Powell, C., Walker, S. & Himes, J. (1994). The Long-Term Follow-up of Severely Malnourished Children Who Participated in an Intervention Program. *Child Development* 65: 428–93.

Grosh, M. & Glewwe, P., eds. (2000). *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study*, vols. 1, 2, e 3. Washington, DC: Banco Mundial.

Grosh, M., del Ninno, C., Tesliuc, E. & Ouerghi, A. (2008). *For Protection and Promotion: The Design and Implementation of Effective Safety Nets*. Washington, DC: Banco Mundial.

Jalan, J. & Ravallion, M. (2003a). Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching. *Journal of Business & Economic Statistics* 21 (1): 19–30.

——. (2003b). Does Piped Water Reduce Diarrhea for Children in Rural India? *Journal of Econometrics* 112 (1): 153–73.

Kimmel, A. (1988). *Ethics and Values in Applied Social Research*. California: Sage Publications.

King, E. & Behrman, J. (2009). Timing and Duration of Exposure in Evaluations of Social Programs. *World Bank Research Observer* 24 (1): 55–82.

King, E., Orazem, P e Paterno, E. (2008). Promotion with and without Learning: Effects on Student Enrollment and Dropout Behavior. *Policy Research Working Paper Series 4722*. Washington, DC: Banco Mundial.

Levy, S. & Rodríguez, E. (2005). *Sin Herencia de Pobreza: El Programa Progresa-Oportunidades de México*. Washington, DC: Banco Interamericano de Desenvolvimento.

NIH (US National Institute of Health). (2006). Regulations and Ethical Guidelines e Belmont Report. Office of Human Subjects Research. Disponível em: http:// ohsr.od.nih. gov/index.html.

Newman, J., Pradhan, M., Rawlings, L., Ridder, G., Coa, R. & Evia, J. (2002). An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund. *World Bank Economic Review* 16 (2): 241–74.

Rosenbaum, P. (2002). Observational Studies. Springer Series in Statistics. Rosenbaum, P., & Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies of Causal Effects. *Biometrika* 70 (1): 41–55.

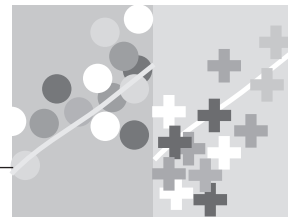
Schultz, P. (2004). School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program. *Journal of Development Economics* 74 (1): 199–250.

Skoufias, E. & McClafferty, B. (2001). Is *Progresa* Working? Summary of the Results of an Evaluation by IFPRI. Washington, DC: Instituto Internacional de Pesquisa sobre as Políticas Alimentares.

USAID (U.S. Agency for International Development). (2008). Procedures for Protection of Human Subjects in Research Supported by USAID. Disponível em: http://www.usaid.gov/policy/ads/200/humansub.pdf.

U.S. Department of Health and Human Services. (2010). International Compilation of Human Research Protections. Office for Human Research Protections. Disponível em: http://www.hhs.gov/ohrp/ international/HSPCompilation.pdf.

Arquivo Nacional dos EUA. (2009). Protection of Human Subjects. *U.S. Code of Federal Regulations*, Título 22, Parte 225.



CAPÍTULO 11

Escolhendo a Amostra

Uma vez escolhido o método para selecionar o grupo de comparação, o próximo passo no planejamento de uma avaliação de impacto é determinar de quais dados você precisa e qual é a amostra necessária para estimar precisamente as diferenças de resultados entre o grupo de tratamento e o grupo de comparação. Você deve determinar o tamanho da amostra e o modo de extrair da população de interesse as unidades da amostra.

De que Tipos de Dados eu Preciso?

São necessários dados de boa qualidade para se avaliar o impacto da intervenção sobre os resultados de interesse. A cadeia de resultados, discutida no Capítulo 2, fornece uma base para definir quais indicadores devem ser medidos e quando. É necessário, antes de tudo, obter dados sobre os indicadores de resultados diretamente afetados pelo programa. No entanto, a avaliação de impacto não deve medir apenas resultados diretamente atribuíveis ao programa. Dados sobre indicadores de resultados que o programa afete indiretamente, ou indicadores que capturem o impacto não intencional do programa, maximizarão o valor da informação que a avaliação de impacto gera. Como discutido no Capítulo 2, os indicadores de resultados devem, preferencialmente, ser escolhidos de modo que sejam "SMART": iniciais, em inglês, dos adjetivos "específicos", "mensuráveis", "atribuíveis",

"realistas" e "direcionados" (*specific, measurable, attributable, realistic* and *targeted*).

Normalmente, as avaliações de impacto são realizadas ao longo de vários períodos; deve-se determinar quando os indicadores de resultados devem ser medidos. Seguindo a cadeia de resultados, pode-se estabelecer uma hierarquia de indicadores de resultados que vão desde indicadores de curto prazo, como a frequência escolar, no contexto de um programa de educação, até os de prazo mais longo, tais como o nível de escolaridade atingido pelos alunos ou os resultados por eles obtidos no mercado de trabalho. Para medir o impacto de forma convincente ao longo do tempo, são necessários dados a começar pela linha de base. A seção sobre programação das avaliações, no Capítulo 10, esclarece quando coletar os dados.

Conceito-chave:
Indicadores são necessários ao longo da cadeia de resultados, para medir resultados finais, resultados intermediários, o fornecimento da intervenção, fatores exógenos e características de controle.

Como veremos, alguns indicadores podem não ser passíveis de avaliação de impacto, no caso de amostras relativamente pequenas. Detectar os impactos de indicadores de resultados extremamente variáveis, eventos raros - ou que tenham a possibilidade de ser afetados apenas marginalmente por determinada intervenção - pode exigir amostras demasiadamente grandes. Identificar o impacto de uma intervenção sobre as taxas de mortalidade materna, por exemplo, será viável apenas em uma amostra que contenha muitas mulheres grávidas. Nesse caso, pode ser aconselhável concentrar a avaliação de impacto em indicadores para os quais haja poder estatístico suficiente para detectar o efeito.

Além de indicadores de resultados, também é útil considerar o seguinte:

- *Dados administrativos sobre o fornecimento da intervenção.* No mínimo, dados de monitoramento que permitam que se saiba quando um programa é iniciado e quem recebe os benefícios, para fornecer uma medida da "intensidade" da intervenção nos casos em que esta pode não ser oferecida a todos os beneficiários com o mesmo conteúdo, qualidade ou duração.

- *Dados sobre fatores exógenos que podem afetar o resultado de interesse.* Esses dados possibilitam o controle das influências externas. Esse aspecto é particularmente importante quando se utilizam métodos de avaliação apoiados em mais pressupostos do que os métodos aleatórios. Levar em consideração esses fatores também ajuda a aumentar o poder estatístico.

- *Dados sobre outras características.* Incluir controles adicionais ou analisar a heterogeneidade dos efeitos do programa em função de certas características possibilita uma estimativa mais precisa dos efeitos do tratamento.

Em suma, são necessários indicadores ao longo de toda a cadeia de resultados, incluindo indicadores de resultados finais, indicadores de resultados intermediários, medidas de fornecimento da intervenção, fatores exógenos e características de controle[1].

O modelo escolhido para a avaliação de impacto também afetará os requisitos dos dados. Por exemplo, se for escolhido o método de pareamento ou o método de diferença-em-diferenças, será necessário coletar dados a partir de um leque muito amplo de características de ambos os grupos, de tratamento e de comparação, para ser possível a realização de uma série de testes de robustez, conforme descrito na Parte 2.

Para cada avaliação, é útil desenvolver uma matriz que relacione as questões de interesse, os indicadores de resultado para cada questão, os outros tipos de indicadores necessários e a fonte de dados, como descrito na Figura 2.3 (Capítulo 2).

Posso usar Dados Existentes?

São quase sempre necessários alguns dados já existentes no começo do programa, para estimar os valores de referência de indicadores ou para realizar cálculos de poder estatístico, como discutiremos mais adiante. Além das fases de planejamento, a disponibilidade de dados já existentes pode diminuir substancialmente o custo da realização de uma avaliação de impacto.

Contudo, dados existentes por si só raramente são suficientes. As avaliações de impacto requerem dados globais, que abranjam uma amostra suficientemente grande para que seja representativa de ambos os grupos, o de tratamento e o de comparação. Raramente se dispõe de dados censitários da população abrangendo inteiramente os grupos de tratamento e de comparação. Mesmo quando esses censos existem, eles podem conter apenas um conjunto limitado de variáveis, ou ser atualizados com pouca frequência. Pesquisas domiciliares nacionalmente representativas podem conter um conjunto abrangente de variáveis de resultados, mas raramente contêm observações suficientes sobre ambos os grupos - o de tratamento e o de comparação - para realizar uma avaliação de impacto. Suponha, por exemplo, que você esteja interessado em avaliar um programa grande, nacional, que atinja 10 por cento dos lares em um determinado país. Se for realizada uma pesquisa nacional representativa com 5.000 domicílios a cada ano, esta poderá conter aproximadamente 500 domicílios que recebem o programa em questão. Essa amostra é grande o suficiente para realizar uma avaliação de impacto? Os cálculos de poder estatístico podem responder a esta pergunta mas, na maioria dos casos, a resposta é não.

Ainda assim, a possibilidade de utilização de *dados administrativos* existentes para realizar avaliações de impacto deve ser seriamente considerada. Dados administrativos são dados coletados por órgãos implementadores de programas, frequentemente no ponto de entrega do serviço, como parte de suas operações regulares. Em alguns casos, os dados de monitoramento contêm indicadores de resultado. Por exemplo, as escolas podem registrar matrículas, frequência, ou resultados dos testes dos alunos e os centros de saúde podem registrar a antropometria e a situação de vacinação ou de saúde dos pacientes. Algumas avaliações retrospectivas influentes têm se embasado em registros administrativos (por exemplo, Galiani, Gertler e Schargrodsky 2005, sobre a política de água na Argentina).

Para determinar se os dados existentes podem ser usados em determinada avaliação de impacto, devem ser consideradas as seguintes perguntas:

- *Tamanho*. Os conjuntos de dados existentes são grandes o bastante para detectar alterações nos indicadores de resultados com suficiente poder estatístico?
- *Amostragem*. Os dados existentes estão disponíveis para tanto o grupo de tratamento quanto o grupo de comparação? As amostras existentes foram tiradas de uma base de amostragem que coincide com a população de interesse? As unidades foram extraídas da base de amostragem a partir de um processo de amostragem probabilística?
- *Escopo*. Os dados existentes contêm todos os indicadores necessários para responder às questões de interesse sobre a política?
- *Frequência*. Os dados existentes são coletados com a frequência adequada? Estão disponíveis para todas as unidades da amostra ao longo do tempo?

Apenas em casos relativamente raros os dados existentes são adequados para avaliações de impacto. Como resultado, muito provavelmente você terá que orçar uma nova coleta de dados. Embora a coleta de dados acarrete, em geral, um custo expressivo, é também um investimento de alto retorno, do qual depende a qualidade da avaliação.

Em alguns casos, os dados necessários para a avaliação de impacto podem ser coletados com a implantação de novos sistemas de informação. Isso deve ser feito de acordo com o desenho da avaliação, de modo que os indicadores de resultados sejam coletados para um grupo de tratamento e de comparação em múltiplas ocasiões. Podem ser necessários novos sistemas de informação antes de novas intervenções, de forma que os centros administrativos usem o novo sistema de informação antes de receber a intervenção a ser avaliada. Como a qualidade de dados administrativos pode variar,

são necessárias auditoria e verificação externa para garantir a confiabilidade da avaliação. Coletar dados de avaliação de impacto por meio de fontes administrativas, ao invés de por meio de pesquisas, pode reduzir drasticamente o custo de uma avaliação, mas nem sempre é algo factível.

Se os dados administrativos não forem suficientes para a avaliação, provavelmente você terá que confiar em *dados de pesquisas*. Além de explorar a possibilidade de utilizar levantamentos pré-existentes, você também deve verificar se está sendo planejado algum novo esforço nacional de coleta de dados, como a pesquisa demográfica e de saúde (DHS, do inglês *Demographic and Health Survey*, ou Pesquisa de Medição de Padrões de Vida). Se estiver sendo planejada uma pesquisa de medição dos indicadores necessários, pode ser possível *sobreamostrar* a população de interesse. Por exemplo, a avaliação do Fundo Social da Nicarágua complementou uma pesquisa para a medição dos padrões de vida com uma amostra extra de beneficiários (Pradhan e Rawlings 2002). Se for planejada uma pesquisa que cubra a população de interesse, é possível que você consiga introduzir uma questão ou uma série de questões como parte dessa pesquisa.

A maioria das avaliações de impacto exige a coleta de dados em pesquisas, incluindo, pelo menos, uma pesquisa de *linha de base* e uma de *seguimento*. Os dados da pesquisa podem ser de vários tipos, dependendo do programa a ser avaliado e da unidade de análise. A maioria das avaliações se apoia em pesquisas individuais ou de domicílio como principal fonte de dados. Aqui, discutimos alguns princípios gerais da coleta de dados em levantamentos. Mesmo que eles se refiram principalmente às pesquisas domiciliares, os mesmos princípios também se aplicam à maioria dos outros tipos de dados de pesquisa[2].

O primeiro passo para decidir se devem ser usados dados existentes ou coletados novos dados de pesquisa será determinar o tamanho necessário da amostra. Se os dados existentes contiverem um número suficiente de observações, pode ser possível usá-los. Caso contrário, terão de ser coletados dados adicionais. Uma vez decidido que é preciso coletar dados de pesquisa para a avaliação, deve-se

- determinar quem coletará os dados,
- desenvolver e aplicar questionários pilotos,
- realizar trabalhos de campo e de controle de qualidade, e
- processar e armazenar os dados,

O restante deste capítulo discutirá como determinar o tamanho necessário da amostra e como extraí-la. As etapas restantes da coleta de dados são tratadas no Capítulo 12. A implementação desses vários passos

é, geralmente, contratada, mas compreender seu escopo seus principais componentes é essencial para que se possa gerir, de forma eficaz, uma avaliação de impacto de qualidade.

Cálculos de Poder Estatístico: qual o Tamanho de Amostra de que eu Preciso?

A primeira etapa na determinação se os dados existentes podem ser utilizados ou na preparação para coletar dados novos para a avaliação será determinar quão grande a amostra terá de ser. Os cálculos associados são chamados de *"cálculos de poder estatístico"*. Discutimos a intuição básica por trás dos cálculos de poder estatístico, focando o caso mais simples — uma avaliação realizada utilizando-se o método de alocação aleatória e supondo que o não cumprimento das regras de participação no programa não seja um problema (o cumprimento das regras pressupõe que todas as unidades alocadas ao grupo de tratamento são tratadas e todas as unidades alocadas ao grupo de comparação não são).

Conceito-chave:
Os cálculos de poder estatístico indicam o tamanho de amostra exigido em uma avaliação para estimar, com precisão, o impacto de um programa (a diferença nos resultados entre os grupos de tratamento e de comparação).

Por que Cálculos de Poder Estatístico?

Os cálculos de poder estatístico indicam o tamanho mínimo da amostra que é necessário para realizar uma avaliação de impacto e responder, de forma convincente, a questão de interesse para as políticas. Em particular, os cálculos de poder estatístico podem ser utilizados para o seguinte:

- Avaliar se os conjuntos de dados existentes são grandes o suficiente para o propósito de realizar uma avaliação de impacto.
- Evitar coletar informações demais, o que pode ser muito caro.
- Evitar coletar dados de menos. Digamos que você esteja estimando um programa que tenha um impacto positivo sobre seus beneficiários. Se a amostra for demasiadamente pequena, você pode não ser capaz de detectar o impacto positivo e pode, assim, concluir que o programa não tem nenhum efeito. Isso, claro, poderia levar à decisão política de se eliminar o programa, o que seria prejudicial para os potenciais beneficiários e para a sociedade.

Os cálculos de poder estatístico oferecem uma indicação da menor amostra (e menor orçamento) com a qual é possível medir o impacto de um programa - isto é, a menor amostra que permita que sejam detectadas diferenças significativas nos resultados entre os grupos de tratamento e de

comparação. Os cálculos de poder estatístico são cruciais para determinar quais programas têm êxito e quais não.

O Impacto do Programa é Diferente de Zero?

A maioria das avaliações de impacto testa uma hipótese simples incorporada à questão: *o programa tem impacto?* Em outras palavras, *O impacto do programa é diferente de zero?* Responder a esta pergunta requer dois passos:

1. Estimar os resultados médios dos grupos de tratamento e de comparação.
2. Avaliar se existe diferença entre o resultado médio do grupo de tratamento e o resultado médio do grupo de comparação.

Estimando os Resultados Médios dos Grupos de Tratamento e de Comparação.

Vamos supor que você esteja interessado em estimar o impacto de um programa de nutrição sobre o peso das crianças de 5 anos de idade. Suponhamos que 100.000 crianças participaram do programa, que 100.000 crianças não participaram e que as crianças escolhidas para participar foram sorteadas aleatoriamente dentre as 200.000 crianças do país. Como primeiro passo, você terá que estimar o peso médio das crianças que participaram e das que não participaram do programa.

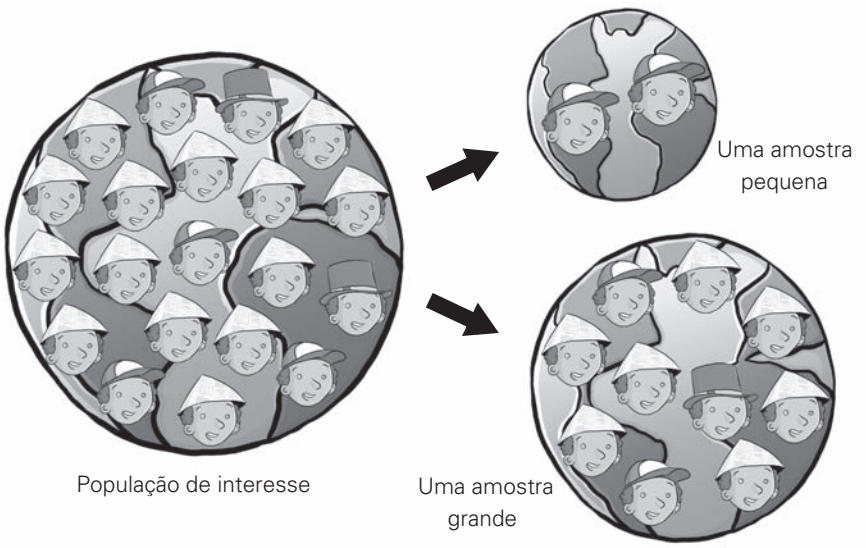
Para determinar o peso médio das crianças participantes[3], pode-se pesar cada uma das 100.000 crianças participantes e, então, tirar a média dos pesos. Claro isso seria extremamente caro de se fazer. Felizmente, não é necessário medir cada criança. A média pode ser calculada usando-se o peso médio de uma amostra extraída da população de crianças participantes[4]. Quanto mais crianças houver na amostra, mais próxima a média da amostra estará da média verdadeira. Quando uma amostra é pequena, o peso médio constitui uma estimativa muito imprecisa da média da população; por exemplo, uma amostra de duas crianças não fornece uma estimativa precisa. Em contrapartida, uma amostra de 10.000 crianças produzirá uma estimativa mais precisa e muito mais próxima do verdadeiro peso médio. Em geral, quanto mais observações houver na amostra, mais confiáveis serão as estatísticas obtidas a partir da amostra[5].

A figura 11.1 ilustra essa intuição. Suponha que você esteja extraindo uma amostra de uma população de interesse — neste caso, as crianças que participaram do programa. Primeiro, você retira uma amostra de apenas duas observações. Isto não garante que a amostra terá as mesmas características

que a população. Pode acontecer de você extrair dois indivíduos com características incomuns. Por exemplo, mesmo que na população de interesse apenas 20 por cento das crianças usem chapéus redondos, você pode facilmente extrair uma amostra de duas crianças que usam chapéus redondos. Obviamente, você não teve sorte ao extrair essa amostra. Extrair amostras maiores diminui as chances de ser azarado. Uma amostra grande é mais provável do que uma amostra pequena de se parecer com a população de interesse. A Figura 11.1 ilustra o que acontece quando se extrai uma amostra grande. Uma amostra grande tem uma propensão muito maior de ter mais ou menos as mesmas características da população: neste exemplo, 20 por cento usam chapéus redondos, 10 por cento usam chapéus quadrados e 70 por cento usam chapéus triangulares.

Portanto, agora sabemos que, com uma amostra maior, teremos uma imagem mais precisa da população de crianças participantes. O mesmo vale para crianças não participantes: à medida que a amostra de crianças não participantes fica maior, saberemos com maior precisão como aquela população se caracteriza. Mas por que devemos nos preocupar? Se formos capazes de estimar o resultado médio (peso) de crianças participantes e não participantes de modo mais preciso, também conseguiremos definir mais

Figura 11.1 Uma Amostra Grande Representará Melhor a População



Fonte: Autores.

precisamente a diferença de peso entre os dois grupos — e esse é o impacto do programa. Colocado de outra forma, se se tem apenas uma vaga ideia do peso médio das crianças nos grupos participante (tratamento) e não participante (comparação), então como se pode ter uma ideia precisa da diferença de peso nos dois grupos? Isso mesmo; não é possível. Na seção seguinte, vamos explorar essa ideia de modo um pouco mais formal.

Comparando os Resultados Médios dos Grupos de Tratamento e de Comparação

Após você ter estimado o resultado médio (peso) do grupo de tratamento (crianças participantes selecionadas por alocação aleatória) e do grupo de comparação (crianças não participantes selecionadas por alocação aleatória), você pode prosseguir e determinar se os dois resultados são diferentes. Esta parte é clara: subtraem-se as médias e verifica-se qual é a diferença. Formalmente, a avaliação do impacto testa a *hipótese nula (ou padrão)*.

H_0: impacto = 0 (A hipótese é de que o programa não tem impacto),

contra a hipótese alternativa:

H_a: impacto $\neq$ 0 (A hipótese é de que o programa tem impacto),

Imagine que, no exemplo do programa de nutrição, você comece com uma amostra de duas crianças tratadas e duas crianças de comparação. Com uma amostra tão pequena, sua estimativa do peso médio das crianças tratadas e das crianças de comparação - bem como sua estimativa da diferença entre os dois grupos - não será muito confiável. Você pode verificar isso extraindo amostras diferentes de duas crianças tratadas e duas de comparação. O que você encontrará é que o impacto estimado do programa varia muito.

Por outro lado, digamos que você comece com uma amostra de 1.000 crianças atendidas e 1.000 crianças de comparação. Como já dissemos, as estimativas do peso médio dos dois grupos serão muito mais precisas. Portanto, a sua estimativa da diferença entre os dois grupos também será mais precisa.

Por exemplo, digamos que você encontre que o peso médio da amostra de crianças de tratamento (participantes) seja de 25,2 quilogramas (kg) e que a média da amostra de crianças de comparação (não participantes) seja de 25 kg. A diferença entre os dois grupos é de 0,2 kg. Se esses números vieram de amostras de duas observações cada, você não terá muita confiança de que o impacto do programa é, realmente, positivo, porque a diferença de 0,2 kg pode ser decorrente da falta de precisão nas estimativas. No entanto,

se esses números vierem de amostras de 1000 observações cada, você terá mais confiança de estar bem perto do impacto real do programa que, neste caso, seria positivo.

A questão-chave torna-se, então, esta: *exatamente quão grande deve ser a amostra para permitir que se saiba que um impacto positivo estimado é devido ao impacto verdadeiro do programa e não à falta de precisão nas estimativas?*

Dois Potenciais Erros nas Avaliações de Impacto

Ao testar se um programa tem impacto, podem ser cometidos dois tipos de erro. O *erro do tipo I* é cometido quando uma avaliação conclui que o programa teve impacto quando, na realidade, não teve impacto nenhum. No caso da hipotética intervenção nutricional, isto aconteceria se, como avaliador, você chegasse à conclusão de que o peso médio das crianças da amostra tratada é mais elevado do que o das crianças na amostra de comparação, embora o peso médio das crianças nas duas populações seja, de fato, igual. Nesse caso, o impacto positivo observado veio puramente da falta de precisão das suas estimativas.

Conceito-chave:
O poder estatístico é a probabilidade de se detectar impacto, caso exista. Uma avaliação de impacto tem poder estatístico mais alto se houver um baixo risco de não detectar impactos reais de programas, isto é, de cometer um erro do tipo II.

Um *erro do tipo II* é o tipo de erro oposto. Um erro do tipo II ocorre quando a avaliação conclui que o programa não teve impacto quando, na realidade, teve impacto. No caso da intervenção nutricional, isto aconteceria se você concluísse que o peso médio das crianças nas duas amostras é o mesmo, mesmo que o peso médio das crianças da população de tratamento seja, de fato, mais elevado do que o das crianças na população de comparação. Mais uma vez, o impacto deveria ter sido positivo, mas por causa da falta de precisão nas estimativas, você concluiu que o programa teve impacto zero.

Ao testar a hipótese de que um programa tenha tido impacto, os estatísticos podem limitar o tamanho de erros do tipo I. Na verdade, a probabilidade de um erro do tipo I pode ser definida por um parâmetro chamado *"nível de confiança"*. O nível de confiança é, muitas vezes, fixado em 5 por cento, significando que se pode ter 95 por cento de confiança em concluir que o programa teve impacto. Se você estiver muito preocupado em cometer um erro do tipo I, poderá definir, conservadoramente, um nível de confiança inferior - por exemplo, de 1 por cento - de modo que tenha 99 por cento de confiança de concluir que o programa teve impacto.

No entanto, os erros do tipo II são também preocupantes para os tomadores de decisão. Muitos fatores afetam a probabilidade de se cometer um erro tipo II, mas o tamanho da amostra é crucial. Se o peso médio de 50.000 crianças atendidas é o mesmo que o peso médio de 50.000 crianças de comparação, então você provavelmente poderá concluir, com confiança, que o

programa não teve impacto. Em contrapartida, se uma amostra de duas crianças de tratamento pesar, em média, o mesmo que a amostra de duas crianças de comparação, é mais difícil chegar-se a uma conclusão segura. O peso médio é semelhante porque a intervenção não teve qualquer impacto ou porque os dados não são suficientes para testar a hipótese em uma amostra tão pequena? Extrair amostras grandes reduz a probabilidade de você só observar crianças que tenham o mesmo peso simplesmente por sorte (ou azar). Em amostras grandes, a diferença de médias entre a amostra tratada e a de comparação proporciona uma estimativa melhor da verdadeira diferença de médias existente entre todas as unidades tratadas e todas as unidades de comparação.

O *poder (ou poder estatístico)* de uma avaliação de impacto é a probabilidade de que ela detectará uma diferença entre os grupos de tratamento e de comparação quando a diferença, de fato, existir. A avaliação de impacto tem um poder estatístico mais alto quando há um baixo risco de não serem detectados os impactos reais dos programas, isto é, de que ocorra um erro tipo II. Os exemplos acima demonstram que o tamanho da amostra é um determinante crucial do poder de uma avaliação de impacto. As seções seguintes irão ilustrar melhor esta questão.

Por que os Cálculos de Poder Estatístico são Cruciais para as Políticas

O propósito de cálculos de poder estatístico é determinar quão grande a amostra precisa ser para evitar a conclusão de que um programa não teve impacto quando, na verdade, teve (erro do tipo II). O poder estatístico de um teste é igual a 1 menos a probabilidade de um erro do tipo II.

Uma avaliação de impacto é de *alto poder* se for improvável a ocorrência de um erro do tipo II, o que significa ser improvável que você se decepcione com resultados que indicam que o programa avaliado não teve impacto quando, na realidade, teve.

Do ponto de vista da política, as *avaliações de impacto de baixo poder estatístico,* com alta probabilidade de erros do tipo II, são não apenas inúteis, mas também muito caras. Uma alta probabilidade de erro do tipo II põe em risco a confiabilidade de resultados negativos da avaliação de impacto. Destinar recursos para as chamadas avaliações de impacto de baixo poder estatístico é, portanto, um investimento arriscado.

Avaliações de impacto de baixo poder estatístico também podem ter consequências práticas dramáticas. Por exemplo, na hipotética intervenção nutricional mencionada anteriormente, se você concluísse que o programa não foi eficaz (mesmo que tenha sido), os tomadores de decisão

provavelmente encerrariam um programa que beneficia as crianças de fato. Assim, é crucial minimizar a probabilidade de erros do tipo II utilizando-se de amostras suficientemente grandes nas avaliações de impacto. É por isso que a realização de cálculos de poder estatístico é tão crucial e relevante.

Cálculos de Poder Estatístico Passo a Passo

Passamos agora aos princípios básicos do cálculo de poder estatístico, focando no caso simples de um programa alocado aleatoriamente. Realizar cálculos de poder estatístico requer o exame das seguintes seis questões:

1. O programa cria *conglomerados*?
2. Qual é o indicador de *resultado*?
3. Você pretende comparar os impactos do programa entre *subgrupos*?
4. Qual é o *nível mínimo de impacto* que justificaria o investimento feito na intervenção?
5. Qual é o *nível de poder estatístico* razoável para a avaliação sendo realizada?
6. Qual é a *média e a variância dos indicadores de resultados* na linha de base?

Cada uma dessas etapas deve estar relacionada ao contexto específico da política em que se decidiu realizar uma avaliação de impacto.

Já mencionamos que o nível mínimo de intervenção de um programa influencia o tamanho necessário da amostra para a avaliação. O primeiro passo em um cálculo de poder estatístico é determinar se o programa que se quer avaliar cria algum *conglomerado*. Uma intervenção cujo nível de intervenção seja diferente do nível em que se gostaria de medir resultados cria um conglomerado. Por exemplo, pode ser necessário implementar um programa em nível de hospital, escola ou município (em outras palavras, por conglomerados), mas se mede o impacto sobre os pacientes, alunos ou moradores (vide Tabela 11.1).[6]

A natureza de qualquer amostra de dados construída a partir de programas que formam conglomerados em sua implementação é um pouco diferente das amostras obtidas de programas que não formam conglomerados. Como resultado, os cálculos de poder estatístico envolverão etapas ligeiramente diferentes, conforme o programa em questão aloque os benefícios aleatoriamente entre os conglomerados ou simplesmente atribua benefícios aleatoriamente entre todas as unidades de uma população. Discutiremos

Tabela 11.1 Exemplos de Conglomerados

Benefício	Nível em que os benefícios são atribuídos (conglomerado)	Unidade na qual o resultado é medido
Transferência condicionada de renda	Município	Domicílio
Tratamento de malária	Escola	Indivíduos
Programa de treinamento	Bairro	Indivíduos

Fonte: Autores.

cada situação por vez. Começaremos com os princípios do cálculo de poder estatístico quando não há conglomerados, ou seja, quando o tratamento é atribuído no nível em que os resultados são observados e, em seguida, passaremos à discussão de cálculos de poder estatístico quando houver conglomerados.

Cálculos de Poder Estatístico sem Conglomerados

Vamos supor que você resolveu a primeira questão, estabelecendo que os benefícios do programa não serão atribuídos por conglomerados. Em outras palavras, o programa a ser avaliado aloca benefícios aleatoriamente entre todas as unidades de uma população elegível. Neste caso, a amostra de avaliação pode ser construída tomando-se uma amostra aleatória simples de toda a *população de interesse*.

As segunda e terceira etapas se referem aos objetivos da avaliação. Na segunda etapa, você deve identificar os *indicadores de resultados* mais importantes para os quais o programa foi concebido. Esses indicadores derivam da questão fundamental da pesquisa de avaliação e do marco conceitual, como discutido na Parte 1. A presente discussão também oferecerá uma visão sobre os tipos de indicadores mais passíveis de serem utilizados em avaliações de impacto.

Terceiro, a principal questão de política da avaliação pode implicar a comparação dos impactos do programa entre *subgrupos*, tais como categorias de renda ou idade. Se for este o caso, então as exigências de tamanho de amostra serão maiores e os cálculos de poder estatístico precisarão ser ajustados. Por exemplo, pode ser que uma questão fundamental de política seja saber se um programa de educação tem um impacto maior sobre alunos do sexo feminino do que sobre alunos do sexo masculino.

Intuitivamente, será necessário um número suficiente de alunos de cada sexo no grupo de tratamento e no grupo de comparação, para detectar um impacto em cada subgrupo. Definir que se quer comparar os impactos do programa entre dois subgrupos pode dobrar o tamanho necessário da amostra. Considerar a heterogeneidade entre vários grupos (por exemplo, por idade) também pode aumentar substancialmente o tamanho necessário da amostra.

Quarto, você deve determinar qual é o nível mínimo de impacto que justificaria o investimento feito na intervenção. Mais do que uma questão técnica, trata-se de uma questão fundamentalmente política. Um programa de transferência condicionada de renda é um investimento que vale a pena caso reduza a pobreza em 5 por cento, 10 por cento ou 15 por cento? Vale a pena implementar um programa ativo de mercado de trabalho se ele aumentar a renda em 5 por cento, 10 por cento ou 15 por cento? A resposta é altamente específica ao contexto mas, em todos os contextos, é necessário determinar a mudança nos indicadores de resultados que justifique o investimento feito no programa. Dito de outra forma, *qual é o nível de impacto abaixo do qual uma intervenção deve ser considerada malsucedida?* A resposta desta pergunta dependerá não apenas do custo do programa e do tipo de benefícios que ele oferece, mas também do custo de oportunidade de não investir em uma intervenção alternativa.

Realizar cálculos de poder estatístico faz com que seja possível ajustar o tamanho da amostra para detectar o efeito mínimo desejado. Para uma avaliação identificar um impacto pequeno, as estimativas de qualquer diferença nos resultados médios entre os grupos de tratamento e de comparação terão que ser muito precisas, exigindo uma amostra grande. Alternativamente, para intervenções consideradas viáveis apenas no caso de acarretarem grandes mudanças nos indicadores de resultados, as amostras necessárias para realizar a avaliação de impacto serão menores. No entanto, o *efeito mínimo detectável* deve ser definido de maneira conservadora, uma vez que qualquer impacto menor do que o efeito mínimo desejado é improvável de ser detectado.

Quinto, o avaliador precisa consultar especialistas em estatística para determinar um *nível de poder estatístico* razoável para a avaliação de impacto planejada. Como mencionado antes, o poder estatístico de um teste é igual a 1 menos a probabilidade de qualquer erro do tipo II. Por conseguinte, o poder estatístico pode variar de 0 a 1, com um valor elevado indicando menor risco de não identificar um impacto existente. O poder estatístico de 80 por cento é uma referência largamente usada para cálculos de poder estatístico. Isso significa que se observará um impacto em 80 por

cento dos casos em que houver ocorrido algum impacto. Um nível de poder estatístico de 0,9 (ou 90 por cento) muitas vezes fornece uma referência útil, mas é mais conservador, aumentando o tamanho necessário das amostras[7].

Sexto, você deve solicitar a um especialista em estatística que estime alguns parâmetros básicos, como a média e a variância de *linha de base dos indicadores de resultados*. Esses valores de referência devem ser, preferencialmente, obtidos a partir de dados existentes, coletados em situação semelhante àquela em que o programa em estudo será implementado[8]. É muito importante notar que quanto maior a variabilidade dos resultados de interesse, mais difícil será estimar um efeito de tratamento confiável. No exemplo hipotético da intervenção nutricional, o peso da criança é o resultado de interesse. Se todos os indivíduos tiverem o mesmo peso na linha de base, será possível calcular o impacto da intervenção nutricional em uma amostra relativamente pequena. Em contrapartida, se os pesos da linha de base entre as crianças variarem muito, será necessária uma amostra muito maior para estimar o impacto do programa.

Uma vez que essas seis etapas tenham sido concluídas, o especialista em estatística pode realizar um cálculo de poder estatístico usando um software estatístico padrão.[9] O cálculo de poder estatístico resultante indicará o tamanho de amostra necessário, dependendo dos parâmetros estabelecidos nas etapas de 1 a 6. Os cálculos propriamente ditos são simples, uma vez que as questões de políticas relevantes tenham sido respondidas (particularmente nas etapas 3 e 4).[10]

Ao buscar o aconselhamento de especialistas em estatística, o avaliador deverá pedir uma análise da *sensibilidade* do cálculo de poder estatístico a alterações nos pressupostos. Ou seja, é importante compreender o quanto o tamanho da amostra terá que aumentar caso os pressupostos sejam mais conservadores (por exemplo, um impacto esperado menor, uma variância maior do indicador de resultado ou um poder estatístico maior). Também é boa prática contratar cálculos de poder estatístico para vários indicadores de resultados, já que o tamanho necessário das amostras pode variar substancialmente se alguns indicadores de resultados variarem mais do que outros.

Finalmente, cálculos de poder estatístico oferecem o tamanho mínimo necessário da amostra. Na prática, os problemas de implementação frequentemente implicam que o tamanho real da amostra seja menor do que o tamanho da amostra planejada. Qualquer diferença deste tipo precisa ser considerada cuidadosamente, mas é aconselhável adicionar uma margem de 10 ou 20 por cento ao tamanho da amostra previsto por cálculos de poder estatístico, para acomodar tais fatores[11].

Conceito-chave:

Os requisitos de amostra aumentam se o efeito mínimo detectável for pequeno, se o indicador de resultado for altamente variável ou um evento raro e se a avaliação pretende comparar impactos entre vários subgrupos.

Qual o Tamanho de Amostra Necessário para que eu Avalie um Programa Ampliado de Subsídio ao Seguro Saúde?

Digamos que o presidente e o ministro da saúde ficaram satisfeitos com a qualidade e os resultados da avaliação do Programa de Subsídio ao Seguro Saúde (HISP), nosso exemplo nos capítulos anteriores. No entanto, antes de intensificar o HISP, eles decidem realizar um piloto com uma versão expandida do programa (que chamam de HISP+). O HISP paga parte do custo do seguro saúde para as famílias pobres das áreas rurais, cobrindo custos de cuidados básicos de saúde e remédios, mas não cobre a internação hospitalar. O presidente e o ministro da saúde se perguntam se um HISP+ expandido, que também abrangesse a hospitalização, diminuiria ainda mais o desembolso com despesas de saúde. Eles pedem a você que desenhe uma avaliação de impacto para avaliar se o HISP+ reduz ainda mais as despesas com saúde das famílias rurais pobres.

Neste caso, a escolha de um desenho de avaliação de impacto não seria um grande desafio para você: o HISP+ tem recursos limitados e não pode ser implementado imediatamente de forma universal. Como resultado, você conclui que a alocação aleatória seria o método mais viável e robusto de avaliação de impacto. O presidente e o ministro da saúde entendem que o método de alocação aleatória funciona bem e o apoiam.

Para finalizar o desenho da avaliação de impacto, você contratou um estatístico que lhe ajudará a definir o tamanho de amostra necessário. Antes que comece a trabalhar, ele lhe pede algumas informações-chave. Para isso, ele usa uma *checklist* com seis questões.

1. O estatístico pergunta se o programa HISP+ gerará conglomerados. Você não tem certeza quanto a isso. Acredita que pode ser possível aleatorizar o pacote de benefícios expandido no nível domiciliar entre todos os domicílios rurais pobres que já se beneficiam do HISP. No entanto, você está ciente de que o presidente e o ministro da saúde podem preferir alocar o programa expandido em nível municipal - e isto criaria conglomerados. O estatístico sugere realizar cálculos de poder estatístico para um caso de referência sem conglomerados e, em seguida, considerar como os resultados mudariam caso existissem conglomerados.

2. O estatístico pergunta qual é o indicador de resultado. Você explica que o governo está interessado em um indicador bem definido: o desembolso das famílias com despesas de saúde. O estatístico procura a fonte de dados mais atualizada para obter valores de referência para esse indicador e sugere a utilização da pesquisa de seguimento da avaliação HISP. Ele observa que, entre as famílias que receberam o HISP, o desembolso anual per capita com despesas de saúde teve um valor médio de $7,84.

3. O estatístico confirma que você não está interessado em medir os impactos do programa para subgrupos, como regiões do país ou subpopulações específicas.

4. O estatístico pergunta qual é o nível mínimo de impacto que justificaria o investimento feito na intervenção. Em outras palavras, que diminuição adicional nos desembolsos com despesas de saúde abaixo da média de referência de $ 7,84 faria esta intervenção valer a pena? Ele ressalta que esta não é uma consideração técnica, mas verdadeiramente uma questão de política. É por isso que um formulador de política como você tem que definir o efeito mínimo que a avaliação deve ser capaz de detectar. Você se lembra de ter ouvido o presidente mencionar que o programa HISP+ seria considerado eficaz se reduzisse os desembolsos das famílias com despesas de saúde em $2. Ainda assim, você sabe que, para efeito da avaliação, talvez seja melhor ser conservador na determinação do impacto mínimo detectável, uma vez que impactos pequenos têm menor probabilidade de serem capturados. Para entender como o tamanho necessário da amostra varia de acordo com o efeito mínimo detectável, você sugere que o estatístico realize cálculos para uma redução mínima de desembolso com despesas de saúde de $1, $2 e $3.

5. O estatístico pergunta qual seria um nível de poder estatístico razoável para a avaliação que está sendo realizada. Ele acrescenta que os cálculos de poder estatístico são, geralmente, realizados para um poder de 0,9, mas se oferece para realizar verificações posteriores de robustez para um nível menos conservador, de 0,8.

6. Finalmente, o estatístico pergunta qual é a variância do indicador de resultado na população de interesse. Ele volta para o conjunto de dados dos domicílios HISP tratados, apontando que o desvio-padrão dos desembolsos com despesas de saúde é de $8.

Equipado com todas essas informações, o estatístico empreende os cálculos de poder estatístico. Conforme acordado, ele começa com o caso mais conservador de um poder estatístico de 0,9. Ele produz os resultados apresentados na Tabela 11.2.

O estatístico conclui que, para detectar uma redução de $2 em desembolsos com despesas de saúde com um poder estatístico de 0,9, a amostra deve conter pelo menos 672 unidades (336 unidades tratadas e 336 unidades de comparação, sem aglomeração). Ele observa que, se você estiver satisfeito em detectar uma redução de $3 em desembolsos com despesas de saúde, uma amostra menor, de pelo menos 300 unidades (150 unidades em cada grupo) seria suficiente. Por outro lado, uma amostra muito maior, de pelo

Tabela 11.2 Tamanho de Amostra Requerido para Vários Efeitos Mínimos Detectáveis (Redução das Despesas das Famílias com Saúde), Poder Estatístico = 0,9, sem Conglomerados

Efeito Mínimo Detectável	Grupo de Tratamento	Grupo de Comparação	Amostra Total
$1	1.344	1.344	2.688
$2	336	336	672
$3	150	150	300

Fonte: Autores.

Nota: O efeito mínimo detectável descreve a redução mínima de desembolsos das famílias com despesas de saúde a ser detectada através da avaliação de impacto.

menos 2.688 unidades (1.344 em cada grupo), seria necessária para detectar uma redução de $1 nos desembolsos com despesas de saúde.

O estatístico produz, então, outra tabela, para um nível de poder estatístico de 0,8. A Tabela 11.3 mostra que os tamanhos necessários de amostra são menores para um poder estatístico de 0,8 do que para um poder estatístico de 0,9. Para detectar uma redução de $2 nos gastos das famílias com saúde, uma amostra total de, pelo menos, 502 unidades seria suficiente. Para detectar uma redução de $3, pelo menos 224 unidades seriam necessárias. No entanto, para detectar uma redução de $1, pelo menos 2008 unidades seriam necessárias na amostra.

O estatístico salienta que os resultados que se seguem são típicos de cálculos de poder estatístico:

- Quanto maior (mais conservador) for o nível de poder estatístico, maior o tamanho necessário da amostra.
- Quanto menor o impacto a ser detectado, maior o tamanho necessário da amostra.

Tabela 11.3 Tamanho de Amostra Requerido para Vários Efeitos Mínimos Detectáveis (Redução das Despesas das Famílias com Saúde), Poder Estatístico = 0,8, sem Conglomerados

Efeito Mínimo Detectável	Grupo de Tratamento	Grupo de Comparação	Amostra Total
$1	1.004	1.004	2.008
$2	251	251	502
$3	112	112	224

Fonte: Autores.

Nota: O efeito mínimo detectável descreve a redução mínima de desembolsos das famílias com despesas de saúde a ser detectada através da avaliação de impacto.

Tabela 11.4 Tamanho de Amostra Requerido para Vários Efeitos Mínimos Detectáveis (Aumento na Taxa de Internação), Poder Estatístico = 0,9, sem Conglomerados

Efeito Mínimo Detectável (ponto percentual)	Grupo de Tratamento	Grupo de Comparação	Amostra Total
1	9.717	9.717	19.434
2	2.430	2.430	4.860
3	1.080	1.080	2.160

Fonte: Autores.

Nota: O efeito mínimo desejado descreve a alteração mínima da taxa de utilização hospitalar (expressa em ponto percentual) a ser detectada pela avaliação de impacto.

O estatístico pergunta se você gostaria de realizar cálculos de poder estatístico para outros resultados de interesse. Você sugere também considerar o tamanho da amostra requerida para detectar se o HISP+ afeta a taxa de internação hospitalar. Na amostra de municípios tratados com HISP, 5% das famílias têm um membro que foi internado em um hospital em um determinado ano. O estatístico gera uma nova tabela, demonstrando que seriam necessárias amostras relativamente grandes para detectar até mesmo alterações maiores na taxa de internação (Tabela 11.4), de 1, 2 ou 3 pontos percentuais da taxa de linha de base de 5%.

A tabela mostra que os requisitos de tamanho de amostra são maiores para esse resultado (a taxa de internação) do que para os desembolsos com despesas de saúde. O estatístico conclui que, se você está interessado em detectar impactos em ambos os resultados, então deve usar as amostras maiores derivadas dos cálculos de poder estatístico realizados nas taxas de internação. Se os tamanhos de amostra dos cálculos de poder estatístico realizados para os desembolsos com saúde forem ser utilizados, então o estatístico sugere informar ao presidente e ao ministro da saúde que a avaliação não terá poder suficiente para identificar efeitos relevantes da política sobre as taxas de internação.

QUESTÃO 8

A. Qual tamanho de amostra você recomendaria para estimar o impacto do HISP+ sobre os desembolsos com despesas de saúde?

B. O tamanho dessa amostra seria suficiente para detectar mudanças na taxa de internação?

Cálculos de Poder Estatístico com Conglomerados

A discussão acima apresentou os princípios da realização de cálculos de poder estatístico para programas que não criam conglomerados. No entanto,

como discutido na Parte 2, alguns programas atribuem benefícios em nível de conglomerado. Vamos agora descrever, brevemente, como os princípios básicos de cálculos de poder estatístico precisam ser adaptados para amostras aglomeradas.

Na presença de conglomerados, um importante princípio orientador é que o número de conglomerados é muito mais importante do que o número de indivíduos dentro dos conglomerados. É necessário um número suficiente de conglomerados para testar, de forma convincente, se um programa teve impacto através da comparação dos resultados em amostras de unidades de tratamento e de comparação.

Se você alocar aleatoriamente o tratamento entre um número pequeno de grupos, os grupos de tratamento e de comparação provavelmente não serão idênticos. A alocação aleatória entre dois distritos, duas escolas ou dois hospitais não garante que os dois grupos sejam semelhantes. Ao contrário, a alocação aleatória de uma intervenção entre 100 distritos, 100 escolas ou 100 hospitais mais provavelmente assegurará que os grupos de tratamento e de comparação serão semelhantes. Em suma, é necessário um número suficiente de conglomerados para assegurar que o equilíbrio seja alcançado. Além disso, o número de conglomerados também é importante para a precisão dos efeitos estimados de tratamento. É necessário um número suficiente de conglomerados para testar, com poder estatístico suficiente, a hipótese de que um programa tem impacto. É, portanto, muito importante assegurar que o número de grupos disponíveis para alocação aleatória seja suficientemente grande.

Seguindo a intuição discutida acima, você pode estabelecer o número de conglomerados necessário para o teste de hipótese através da realização de cálculos de poder estatístico. Realizar cálculos de poder estatístico para as amostras de conglomerados requer um passo extra, além do procedimento básico:

1. O programa cria conglomerados?
2. Qual é o indicador de resultado?
3. Você pretende comparar os impactos do programa entre subgrupos?
4. Qual é o nível mínimo de impacto que justificaria o investimento feito no programa?
5. Qual é a média e a variância do indicador de resultado da linha de base?
6. Qual é a variabilidade do indicador de resultado na população de interesse?
7. Qual é a variabilidade do indicador de resultado dentro dos conglomerados?

Em comparação aos cálculos de poder estatístico sem conglomerado, apenas o último passo é novo: agora você também tem de perguntar a seu especialista em estatística qual o grau de correlação entre os resultados nos conglomerados. No extremo, todos os resultados dentro de um conglomerado são perfeitamente correlacionados. Por exemplo, pode ser que a renda familiar não seja especialmente variável dentro dos municípios, mas que ocorram desigualdades significativas de renda entre municípios. Nesse caso, se você considerar a adição de um indivíduo à sua amostra de avaliação, então adicionar um indivíduo de um novo município proporcionará muito mais poder estatístico adicional do que a adição de um indivíduo de um município está representado.

De fato, neste caso, o segundo morador pode parecer muito semelhante ao morador original já incluído. Em geral, a maior correlação de *intraconglomerados* dos resultados aumenta o número de grupos necessários para atingir um dado nível de poder estatístico.

Conceito-chave:
O número de conglomerados importa muito mais para os cálculos de poder estatístico do que o número de indivíduos nos conglomerados. Pelo menos 30 conglomerados são necessários, frequentemente, em cada um dos grupos de tratamento e de comparação.

Em amostras de conglomerado, cálculos de poder estatístico destacam as decisões entre adicionar grupos e adicionar observações em conglomerados. O aumento relativo do poder estatístico ao se adicionar uma unidade a um novo conglomerado é quase sempre maior do que adicionar uma unidade a um conglomerado existente. Embora o ganho de poder estatístico ao se adicionar um novo conglomerado possa ser dramático, acrescentar conglomerados também pode ter implicações operacionais e aumentar o custo da coleta de dados. A próxima seção mostra como realizar cálculos de poder estatístico com conglomerados no caso de HISP+ e discute algumas das decisões envolvidas.

Em muitos casos, pelo menos 30 a 50 conglomerados em cada grupo de tratamento e de comparação serão necessários para que se obtenha poder estatístico suficiente e garantia de equilíbrio das características da linha de base quando se utilizam métodos de alocação aleatória. No entanto, o número pode variar de acordo com vários parâmetros discutidos acima, bem como o grau de correlação de intraconglomerados. Além disso, o número provavelmente aumentará no caso de se utilizar outros métodos diferentes da alocação aleatória (assumindo que todo o resto seja definido como constante).

Qual o Tamanho de Amostra Necessário para se Avaliar um Programa Ampliado de Subsídio ao Seguro Saúde com Conglomerados?

Após sua primeira conversa com o estatístico sobre cálculos de poder estatístico para o HISP+, você decidiu falar brevemente com o presidente e o

ministro da saúde sobre as implicações de alocar aleatoriamente os benefícios expandidos do HISP+ entre todos os indivíduos da população que recebe o plano HISP básico. Essa consulta revelou que tal procedimento não seria politicamente viável: seria difícil explicar por que uma pessoa receberia benefícios expandidos e seu vizinho, não.

Em vez da aleatorização em nível individual, você sugere selecionar aleatoriamente um número de municípios no HISP para realizar um piloto do HISP+. Assim, todos os moradores dos municípios selecionados seriam elegíveis. Esse procedimento criará conglomerados e, portanto, exigirá novos cálculos de poder estatístico. Agora você quer determinar o tamanho de amostra necessário para avaliar o impacto do HISP+ quando for alocado aleatoriamente por conglomerado.

Você consulta o seu estatístico novamente. Ele lhe garante: só é necessário um pouco de trabalho adicional. Apenas uma pergunta fica sem resposta em sua lista. Ele precisa saber qual é a variabilidade do indicador de resultado nos conglomerados. Felizmente, essa é uma questão que ele também pode responder, usando os dados de seguimento do HISP, a partir dos quais ele descobre que a correlação das despesas de saúde nos municípios é igual a 0,04.

Ele também pergunta se foi imposto um limite superior de número de municípios em que seria viável implementar o novo piloto. Uma vez que o programa agora tem 100 municípios HISP, você explica que é possível ter, no máximo, 50 municípios de tratamento e 50 municípios de comparação no HISP+. Com essa informação, o estatístico produz os cálculos de poder estatístico exibidos na Tabela 11.5 para o valor de 0,9.

O estatístico conclui que, para detectar uma redução de $2 nos desembolsos com despesas de saúde, a amostra tem de conter pelo menos 900 unidades, isto é, 9 unidades por conglomerado em 100 conglomerados. Ele observa

Tabela 11.5 Tamanho de Amostra Requerido para Vários Efeitos Mínimos Detectáveis (Redução das Despesas das Famílias com Saúde), Poder Estatístico = 0,9, Máximo de 100 Conglomerados

Efeito mínimo detectável	Número de conglomerados	Unidades por conglomerado	Amostra total com conglomerados	Amostra total sem conglomerados
$1	Não é viável	Não é viável	Não é viável	2.688
$2	100	9	900	672
$3	85	4	340	300

Fonte: Autores.

Nota: O efeito mínimo desejado descreve a redução mínima de desembolsos domiciliares com despesas de saúde a ser detectada através da avaliação de impacto.

Tabela 11.6 Tamanho de Amostra Requerido para Vários Efeitos Mínimos Detectáveis (Redução das Despesas das Famílias com Saúde), Poder Estatístico = 0,8, Máximo de 100 Conglomerados

Efeito mínimo detectável	Número de conglomerados	Unidades por por conglomerado	Amostra total com conglomerados	Amostra total sem conglomerados
$1	100	102	10.200	2.008
$2	90	7	630	502
$3	82	3	246	224

Fonte: Autores.

Nota: O efeito mínimo detectável descreve a redução mínima de desembolsos domiciliares com despesas de saúde a ser detectada através da avaliação de impacto.

que esse número é maior do que na amostra sob alocação aleatória em nível domiciliar, o que exigiu apenas um total de 672 unidades. Para detectar uma redução de $3 nos desembolsos com despesas de saúde, a amostra teria de incluir, pelo menos, 340 unidades, ou 4 em cada um dos 85 grupos.

No entanto, quando o estatístico tenta estabelecer a amostra necessária para detectar uma redução de $1 nos desembolsos com despesas de saúde, descobre que não seria possível detectar tal efeito com 100 conglomerados. Seriam necessários pelo menos 109 conglomerados e, mesmo assim, o número de observações dentro de cada grupo seria extremamente elevado. Como ele observa, esse resultado destaca que é necessário um grande número de conglomerados para uma avaliação ter poder estatístico suficiente para detectar impactos relativamente pequenos, independentemente do número de observações dentro dos conglomerados.

O estatístico sugere, então, considerar como esses números variarão com um poder estatístico de apenas 0,8 (tabela 11.6). Os tamanhos necessários das amostras, novamente, são menores para um poder estatístico de 0,8 do que para um poder estatístico de 0,9, mas ainda são maiores para a amostra em conglomerado do que para a amostra aleatória simples.

O estatístico, então, lhe mostra como o número total de observações necessárias na amostra varia com o número total de conglomerados. Ele decide repetir os cálculos para um efeito mínimo detectável de $2 e um poder estatístico de 0,9. O tamanho da amostra total necessário para estimar tal efeito aumenta grandemente quando o número de grupos diminui (Tabela 11.7). Com 100 conglomerados, foi necessária uma amostra com 900 observações. Se apenas 30 conglomerados existissem, o total da amostra deveria conter 6.690 observações. Por outro lado, se houvesse 157 conglomerados, seriam necessárias apenas 785 observações.

Tabela 11.7 Tamanho de Amostra Requerido para Detectar um Impacto Mínimo de $2 para Vários Números de Conglomerados, Poder Estatístico = 0,9

Efeito mínimo detectável	Número de conglomerados	Unidades por conglomerado	Amostra total com conglomerados
$2	30	223	6.690
$2	60	20	1.200
$2	86	11	946
$2	100	9	900
$2	120	7	840
$2	135	6	810
$2	157	5	785

Fonte: Autores.

QUESTÃO 9

A. Que tamanho de amostra total você recomendaria para estimar o impacto do HISP+ sobre os desembolsos com despesas de saúde?

B. Em quantos municípios você aconselharia ao presidente e ao ministro da saúde implantar o HISP+?

Em Resumo

Para resumir, a qualidade de uma avaliação de impacto depende diretamente da qualidade dos dados em que se baseia. A esse respeito, amostras construídas apropriadamente e de tamanhos adequados são absolutamente cruciais. Revimos os princípios básicos da realização de cálculo de poder estatístico. Quando realizados no planejamento de uma avaliação, os cálculos de poder estatístico são uma ferramenta essencial para conter os custos da coleta de dados, evitando que se coletem mais dados do que o necessário, enquanto também minimizam o risco de se chegar à conclusão cara e errônea de que um programa não teve impacto porque pouca informação foi coletada. Embora cálculos de poder estatístico requeiram fundamentos técnicos e estatísticos, também exigem uma clara fundamentação da política. Em geral, o aumento do tamanho da amostra produz retornos decrescentes, de maneira que a determinação de uma amostra adequada frequentemente exigirá equilibrar a necessidade de estimativas de impacto precisas com considerações orçamentárias.

Nós nos concentramos no caso de referência de uma avaliação de impacto implementada usando o método de alocação aleatória. Este é o cenário mais simples e, portanto, o mais adequado para transmitir a intuição por trás dos cálculos de poder estatístico. Ainda assim, muitos aspectos práticos de

nossos cálculos de poder estatístico não foram discutidos e desvios dos casos básicos, discutidos aqui, precisam ser considerados com cuidado. Por exemplo, métodos quase-experimentais de avaliação de impacto exigem, em geral, amostras maiores do que a alocação aleatória. As exigências de tamanho da amostra também aumentam se o risco de viés estiver presente nos efeitos de tratamento estimados ou quando o cumprimento das regras de participação for imperfeito. Essas questões estão fora do escopo deste livro, mas são discutidas em Spybrook et al. (2008) e Rosenbaum (2009, capítulo 14) com mais detalhes. Existem inúmeras ferramentas para aqueles interessados em explorar desenhos amostrais mais profundamente. Por exemplo, a Fundação W.T. Grant desenvolveu o software livre *Optimal Design Software for Multi-Level and Longitudinal Research*, que é útil na análise de poder estatístico na presença de conglomerados. Na prática, muitas agências que encomendam avaliações contratam um especialista para realizar cálculos de poder estatístico e o especialista deve ser capaz de aconselhar nos casos em que métodos diferentes da alocação aleatória são utilizados.

Conceito-chave:
métodos quase-experimentais de avaliação de impacto exigem, em geral, amostras maiores do que a alocação aleatória.

Decidindo sobre a Estratégia de Amostragem

Tamanho não é o único fator relevante na garantia de que uma amostra seja adequada para uma avaliação de impacto. O processo pelo qual uma amostra é extraída da população de interesse também é crucial. Os princípios de amostragem podem ser guias para a extração de amostras representativas. A amostragem requer três passos:

1. Determinar a *população de interesse*.
2. Identificar uma *base de amostragem*.
3. *Extrair* tantas unidades da base de amostragem quanto exigidas pelos cálculos de poder estatístico.

Primeiro, *a população de interesse* tem de ser definida muito claramente[12]. Fazer isso requer precisão na definição da unidade de observação para a qual os resultados serão medidos, com especificação clara da cobertura geográfica ou quaisquer outros atributos relevantes que caracterizem a população. Por exemplo, caso se esteja gerindo um programa de desenvolvimento da primeira infância, poderá haver interesse em medir os resultados cognitivos de crianças entre as idades de 3 e 6 em todo o país apenas dentre as crianças de áreas rurais, ou somente dentre as crianças matriculadas na educação infantil.

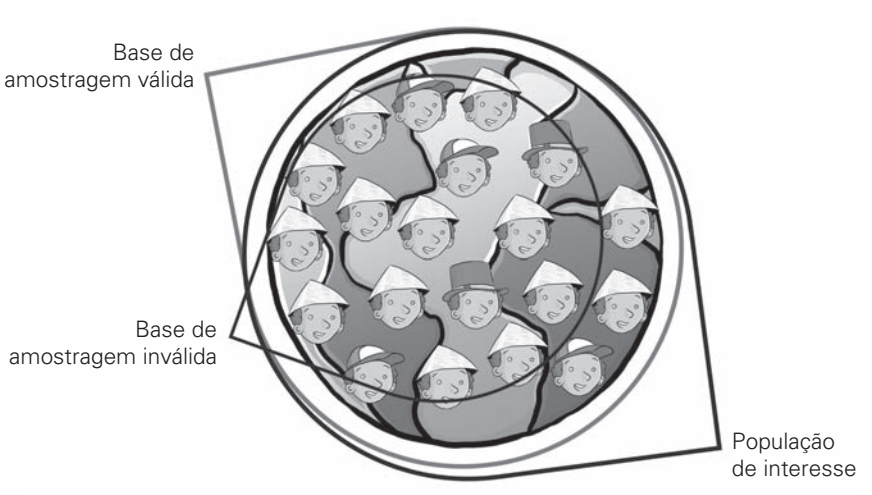
Segundo, uma vez que a população de interesse tenha sido definida, deverá ser estabelecida uma *base de amostragem*. A base de amostragem é a lista mais abrangente que pode ser obtida de unidades na população de interesse. Idealmente, a base de amostragem deve coincidir exatamente com a população de interesse. Por exemplo, um censo completo e totalmente atualizado da população de interesse constituiria uma base de amostragem ideal. Na prática, as listas já existentes - como os censos populacionais, censos de instalações ou listas de inscrição - são frequentemente utilizadas como bases de amostragem.

Conceito-chave:
Uma base de amostragem é a lista mais abrangente que pode ser obtida de unidades na população de interesse. O viés de cobertura ocorre se a base de amostragem não coincidir perfeitamente com a população de interesse.

É necessária uma base de amostragem adequada para garantir que as conclusões obtidas a partir da análise de uma amostra possam ser generalizadas para toda a população. De fato, uma base de amostragem que não coincide exatamente com a população de interesse cria um *viés de cobertura*, como ilustrado na Figura 11.2. Na presença de viés de cobertura, os resultados da amostra não têm total validade externa para o conjunto da população de interesse, apenas para a população incluída na base de amostragem. Como resultado, coberturas enviesadas obscurecem a interpretação dos resultados da avaliação de impacto, uma vez que não fica claro de qual população são provenientes.

Ao considerar a extração de uma nova amostra ou a avaliação da qualidade de uma amostra existente, é importante determinar se a melhor base

Figura 11.2 Uma Base de Amostragem Válida Cobre Toda a População de Interesse



Fonte: Autores.

de amostragem disponível coincide com a população de interesse. O grau em que as estatísticas calculadas a partir da amostra podem ser generalizadas para a população de interesse como um todo depende da magnitude do viés de cobertura - em outras palavras, da falta de coincidência entre a base de amostragem e a população de interesse.

O viés de cobertura pode ocorrer, por exemplo, caso você esteja interessado em todos os domicílios de um país e use uma lista telefônica como base de amostragem, de modo que qualquer domicílio sem telefone não será selecionado para a amostra. Isso pode enviesar os resultados da avaliação, caso os domicílios sem telefone também tenham outras características diferentes daquelas da população de interesse e, ainda, se essas características afetarem o modo como os domicílios se beneficiam da intervenção. Por exemplo, os domicílios sem telefone podem estar em áreas rurais remotas. Se você estiver interessado em avaliar o impacto de um programa de formação profissional, omitir a população mais isolada afetará os resultados da avaliação, pois esses domicílios tendem a ter mais dificuldade em acessar o mercado de trabalho.

Vieses de cobertura constituem um risco real e a construção de bases de amostragem requer um esforço cuidadoso. Por exemplo, os dados do censo podem conter a lista de todas as unidades em uma população. No entanto, se tiver passado muito tempo entre o censo e o momento em que os dados da amostra foram coletados, a base de amostragem pode já não estar totalmente atualizada, criando um viés de cobertura. Além disso, os dados censitários podem não conter informações suficientes sobre atributos específicos para construir uma base de amostragem. Se a população de interesse consistir de crianças que frequentam a educação infantil e o censo não contiver dados sobre matrícula pré-escolar, serão necessários dados complementares de matrícula ou listas de instalações[13].

Uma vez que você tenha identificado a população de interesse e uma base de amostragem, o próximo passo é a escolha de um método para extrair a amostra. Vários procedimentos alternativos podem ser usados. Os métodos de *amostragem probabilística* são os mais rigorosos, pois atribuem uma probabilidade bem definida a cada unidade a ser extraída. Os três principais métodos de amostragem probabilística são os seguintes[14]:

Conceito-chave:

Amostragem é o processo através do qual as unidades são extraídas de uma base de amostragem. Amostragem probabilística atribui uma probabilidade definida para cada unidade a ser extraída.

- *Amostragem aleatória*. Cada unidade da população tem exatamente a mesma probabilidade de ser extraída[15].
- *Amostragem aleatória estratificada*. A população é dividida em grupos (por exemplo, masculino e feminino) e a amostragem aleatória é realizada com cada grupo. Consequentemente, cada unidade em cada grupo

(ou estrato) tem a mesma probabilidade de ser extraída. Desde que cada um dos grupos seja suficientemente grande, a amostragem estratificada possibilita extrair inferências sobre resultados, não apenas no nível da população, mas também com respeito a cada grupo. A estratificação é essencial para avaliações que visam comparar os impactos do programa entre os subgrupos.

- *Amostragem por Conglomerados.* As unidades são agrupadas em conglomerados e é extraída uma amostra aleatória de conglomerados, de forma que ou todas as unidades nesses conglomerados constituem a amostra, ou um número de unidades de cada conglomerado é extraído aleatoriamente. Isto significa que cada conglomerado tem uma probabilidade bem definida de ser selecionado e as unidades dentro de um conglomerado selecionado também têm uma probabilidade bem definida de serem extraídas.

No contexto de uma avaliação de impacto, o procedimento para extrair uma amostra frequentemente deriva das regras de elegibilidade do programa em avaliação. Como descrito na discussão sobre o tamanho da amostra, se a menor unidade viável de implementação for maior do que a unidade de observação, a alocação aleatória de benefícios criará conglomerados. Por esse motivo, a amostragem por conglomerados muitas vezes surge em estudos de avaliação de impacto.

A amostragem não probabilística pode criar erros sérios de amostragem. Às vezes, se usa a *amostragem intencional* ou a *amostragem de conveniência* em vez dos procedimentos bem definidos de amostragem probabilística discutidos acima. Nesses casos, erros de amostragem podem ocorrer mesmo que a base de amostragem capture toda a população e não haja viés de cobertura. Para ilustrar, suponha que uma pesquisa nacional seja realizada pedindo a um grupo de entrevistadores que colete dados domiciliares da residência mais próxima da escola em cada município. Quando é usado um procedimento desses, de amostragem não probabilística, é provável que a amostra não seja representativa da população de interesse como um todo. Em particular, surgirá um viés de cobertura, uma vez que residências remotas não serão pesquisadas.

No final, é necessário prestar bastante atenção à base de amostragem e ao procedimento de amostragem para determinar se os resultados obtidos a partir de determinada amostra têm validade externa para toda a população de interesse. Mesmo se a base de amostragem tiver uma cobertura perfeita e for usado um procedimento de amostragem probabilística, erros não amostrais também podem limitar a validade externa da amostra. Discutiremos erros não amostrais no próximo capítulo.

Notas

1. Dados de custos também são necessários para análise custo-benefício.
2. Para uma referência detalhada sobre pesquisas domiciliares, vide Grosh e Glewwe (2000) e ONU (2005). Dal Poz e Gupta (2009) discutem algumas questões específicas à coleta de dados no setor de saúde.
3. Neste ponto, a discussão pode ser aplicada a qualquer população — toda a população de interesse, a população de tratamento ou a população de comparação.
4. Neste contexto, o termo "população" não se refere à população do país, mas a todo o grupo de crianças de interesse, a "população de interesse".
5. Esta intuição é formalizada por um teorema chamado de "teorema central do limite". Formalmente, para um resultado y, o teorema do limite central afirma que a média $\bar{y}$ da amostra, em média, constitui uma estimativa válida da média da população. Além disso, para uma amostra de tamanho n e para uma variância da população σ^2, a variância da média da amostra é inversamente proporcional ao tamanho da amostra:

   $$var(\bar{y}) = \frac{\sigma^2}{n}$$

   À medida que o tamanho da amostra *n* aumenta, a variância das estimativas obtidas da amostra tende para 0. Em outras palavras, a média é estimada de modo mais preciso em amostras grandes do que em amostras pequenas.
6. A atribuição dos benefícios por conglomerado é frequentemente necessária por considerações sociais ou políticas que tornam a aleatorização dentro de conglomerados impossível. No contexto de uma avaliação de impacto, se torna necessário aglomerar muitas vezes, por causa de prováveis efeitos de transbordamento ou por contágio dos benefícios do programa entre indivíduos nos conglomerados.
7. Juntamente com o poder estatístico, também precisa ser definido um nível de confiança que fixe uma probabilidade aceitável de erro tipo I, normalmente em 0,05 (ou 0,01, no caso de um nível conservador).
8. Ao calcular o poder estatístico a partir de uma linha de base, a correlação entre os resultados ao longo do tempo também deve ser levada em conta nos cálculos.
9. Por exemplo, Spybrook et al. (2008) apresentaram o Optimal Design, um software de fácil utilização para realizar cálculos de poder estatístico.
10. Geralmente, é desejável que os grupos de tratamento e de comparação sejam do mesmo tamanho. De fato, para um determinado número de observações em uma amostra, o poder estatístico é maximizado atribuindo-se metade das observações ao grupo de tratamento e metade ao grupo de comparação. No entanto, os grupos de tratamento e de comparação nem sempre têm de ser do mesmo tamanho. Informe o estatístico sobre quaisquer restrições quanto a ter dois grupos de igual tamanho ou qualquer razão para ter dois grupos de tamanhos diferentes.
11. O capítulo 12 discutirá as questões de não resposta e de atrição em mais detalhes.

12. No contexto de uma avaliação de programa, a população total de interesse pode ser atribuída ao grupo de tratamento ou ao grupo de comparação. Esta seção discute, em termos gerais, como extrair uma amostra da população total de interesse.
13. Se for utilizada a amostragem de conglomerado e a lista de unidades dentro dos conglomerados estiver desatualizada, você deve considerar a possibilidade de realizar uma enumeração completa das unidades dentro de cada conglomerado. Por exemplo, se um município for selecionado para a amostra, a agência responsável pela coleta de dados poderia começar listando todos os domicílios do município antes de realizar a pesquisa propriamente dita.
14. Vide Cochran (1977); Lohr (1999); Kish (1995); Thompson (2002) ou, em um nível mais básico, Kalton (1983), para uma discussão detalhada sobre amostragem (incluindo outros métodos, como amostragem sistemática ou amostragem de multiestágios) para além dos conceitos básicos discutidos aqui. Grosh e Muñoz (1996); Fink (2008); Iarossi (2006) e ONU (2005), todos oferecem orientações práticas para a amostragem.
15. Estritamente falando, as amostras são extraídas de bases de amostragem. Em nossa discussão, supomos que a base de amostragem coincida perfeitamente com a população.

Referências

Cochran, W. (1977). Sampling Techniques. 3ª ed. Nova York: John Wiley.

Dal Poz, M. & Gupta, N. (2009). Assessment of Human Resources for Health Using Cross-National Comparison of Facility Surveys in Six Countries. Human Resources for Health 7: 22.

Fink, A. (2008). *How to Conduct Surveys: A Step by Step Guide.* 4ª ed. Beverly Hills, CA: Sage Publications.

Galiani, S., Gertler, P. & Schargrodsky, E. (2005). Water for Life: The Impact of the Privatization of Water Services on Child Mortality. *Journal of Political Economy* 113(1): 83–120.

Grosh, M. & Glewwe, P., eds. (2000). *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study.* Washington, DC: Banco Mundial.

Grosh, M. & Muñoz. J. (1996). A Manual for Planning and Implementing the Living Standards Measurement Study Survey. *LSMS Working Paper 126*. Washington, DC: Banco Mundial.

Iarossi, G. (2006). *The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents.* Washington, DC: Banco Mundial.

Kalton, G. (1983). *Introduction to Survey Sampling.* Beverly Hills, CA: Sage Publications.

Kish, L. (1995). *Survey Sampling.* Nova York: John Wiley.

Lohr, S. (1999). Sampling: Design and Analysis. Pacific Grove, CA: Brooks Cole.

ONU (Nações Unidas). 2005. *Household Sample Surveys in Developing and Transition Countries.* Nova York: Nações Unidas.

Pradhan, M. & Rawlings, L. (2002). The Impact and Targeting of Social Infrastructure Investments: Lessons from the Nicaraguan Social Fund. *World Bank Economic Review* 16 (2): 275–95.

Rosenbaum, P. (2009). *Design of Observational Studies.* Nova York: Springer Series in Statistics.

Spybrook, J., Raudenbush, S., Liu, X., Congdon, R. & Martinez, A. (2008). *Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software.* Nova York: William T. Grant Foundation.

Thompson, S. (2002). *Sampling.* 2nd ed. Nova York: John Wiley.

CAPÍTULO 12

Coletando Dados

No capítulo 11, discutimos os tipos de dados necessários para uma avaliação e observamos que a maioria das avaliações requer a coleta de novos dados. Em seguida, discutimos como determinar o tamanho da amostra necessário e como extrair a amostra. Neste capítulo, examinaremos os passos da coleta de dados. Um claro entendimento dessas medidas ajudará você a garantir que a avaliação de impacto se baseie em dados de qualidade, que não comprometam o projeto de avaliação. Como primeiro passo, você precisará contratar a ajuda de uma empresa ou órgão do governo especializado em coleta de dados. Paralelamente, você contratará o desenvolvimento de um questionário apropriado. A entidade de coleta de dados recrutará e treinará o pessoal de campo e realizará um teste piloto do questionário. Depois de fazer os ajustes necessários, a empresa ou entidade estará apto a prosseguir com o trabalho de campo. Finalmente, os dados coletados terão de ser digitalizados ou processados e validados, antes de serem entregues e utilizados.

Contratando Ajuda para Coletar Dados

Você precisará designar o órgão responsável pela coleta de dados logo no início. Alguns *trade-offs* importantes devem ser considerados ao decidir

quem deve coletar dados de avaliação de impacto. Os candidatos potenciais para o trabalho abrangem

- a instituição responsável pela execução do programa;
- outra instituição do governo com experiência em coleta de dados (como o órgão local de estatística); ou
- uma empresa independente ou um grupo de pesquisa especializado em coleta de dados.

A entidade de coleta de dados precisa sempre trabalhar em estreita coordenação com o *órgão de execução do programa*. Como os dados de linha de base precisam ser coletados antes do início de qualquer operação do programa, é necessária estreita coordenação para assegurar que nenhuma operação do programa seja executada antes da coleta de dados ser feita. Quando dados da linha de base forem necessários para o funcionamento do programa (por exemplo, dados de um índice de focalização, no âmbito de uma avaliação baseada em um desenho de regressão descontínua), a entidade responsável pela coleta de dados deve ter a capacidade de, rapidamente, processá-los e transferi-los para a instituição encarregada das operações do programa. Também é necessária estreita coordenação na programação da coleta de dados de *seguimento*. Se for escolhida uma implementação com seleção aleatória, por exemplo, a pesquisa de seguimento deverá ser realizada antes do programa ser estendido para o grupo de comparação, para que se evite contaminação.

Um fator extremamente importante na decisão de quem deve coletar os dados é que devem ser cumpridos os mesmos procedimentos de coleta de dados para ambos os grupos: o de comparação e o de tratamento. Muitas vezes, o órgão de execução tem contato apenas com o grupo de tratamento e, assim, não está bem posicionado para coletar dados referentes aos grupos de comparação. Porém, utilizar diferentes órgãos de coleta de dados para os grupos de tratamento e de comparação é arriscado, pois pode gerar diferenças nos resultados mensurados nos dois grupos, pelo simples fato dos procedimentos de coleta de dados diferirem. Se o órgão de execução não puder coletar dados de forma eficaz para ambos os grupos (de tratamento e de comparação) deve-se considerar seriamente a possibilidade de contratar-se um parceiro para fazê-lo.

Em alguns contextos, pode também ser aconselhável que o serviço de coleta de dados seja realizado por um órgão independente, para garantir a objetividade. A preocupação de que o órgão executor do programa não colete dados objetivos pode não se justificar, mas um organismo de coleta de dados independente, que não tenha interesse na

avaliação dos resultados, pode agregar credibilidade ao esforço geral da avaliação de impacto.

Como a coleta de dados envolve uma sequência complexa de operações, recomenda-se que uma entidade especializada e experiente seja responsável por ela. Poucos órgãos de implementação de programas têm experiência suficiente para coletar os dados de larga escala e de alta qualidade necessários para uma avaliação de impacto. Na maioria dos casos, você terá que considerar a contratação de uma instituição local, tal como o órgão nacional de estatística, uma empresa especializada ou um grupo de pesquisa (*think tank*).

Contratar uma instituição local, como o órgão *nacional de estatística,* pode dar à instituição exposição a estudos de avaliação de impacto e ajudá-la a reforçar suas capacidades. No entanto, órgãos locais de estatísticas nem sempre têm a capacidade de assumir incumbências extras, além de suas atividades regulares. Também lhes pode faltar a experiência necessária em pesquisa de campo para avaliações de impacto - por exemplo, experiência exitosa em rastrear indivíduos ao longo do tempo. Se tais restrições existirem, a contratação de uma *empresa independente* ou *think tank* especializado em coleta de dados pode ser mais prática.

Não é necessário usar a mesma entidade para coletar informações na linha de base e nas pesquisas de seguimento. Por exemplo, na avaliação do impacto de um programa de capacitação, onde a população de interesse compreende os indivíduos que se inscreveram no curso, a instituição responsável pelo curso poderia coletar os dados da linha de base quando os indivíduos se matriculassem. É improvável, no entanto, que o mesmo órgão também seja a melhor escolha para coletar as informações de seguimento para os grupos de tratamento e de comparação. Nesse contexto, contratar rodadas de coleta de dados separadamente tem suas vantagens, mas devem ser feitos esforços para não se perder nenhuma informação, entre as rodadas, que possa ser útil para rastrear domicílios ou indivíduos, bem como garantir que dados de linha de base e de seguimento sejam medidos de maneira consistente.

Para determinar a melhor instituição para a coleta de dados de avaliação de impacto, todos estes fatores — experiência na coleta de dados, capacidade de coordenar com o órgão de implementação do programa, independência, oportunidades de construção de capacidade, adaptabilidade ao contexto da avaliação do impacto — devem ser pesados, juntamente com a qualidade provável dos dados obtidos em cada caso. Uma maneira eficaz de identificar a organização melhor posicionada para coletar os dados de qualidade é redigir termos de referência e pedir que as organizações enviem propostas técnicas e financeiras.

Como a pronta entrega e a qualidade dos dados é crucial para a confiabilidade da avaliação de impacto, o contrato com o órgão responsável pela coleta de dados deve ser estruturado com cuidado. O escopo do trabalho e os resultados esperados devem ser delineados de forma extremamente clara. Além do mais, muitas vezes é aconselhável introduzir incentivos em contratos e vincular esses incentivos a indicadores claros de qualidade dos dados. Por exemplo, como iremos salientar abaixo, a taxa de não resposta é um indicador chave de qualidade dos dados. Para criar incentivos para que os órgãos de coleta de dados minimizem a ausência de resposta, o contrato pode estipular um custo unitário para os primeiros 90% da amostra, um custo unitário maior para as unidades entre 90% e 95% e, novamente, um custo unitário mais alto para as unidades entre 95% e 100%. Alternativamente, pode ser confeccionado um contrato separado para a empresa de pesquisa rastrear os não respondentes.

Desenvolvendo o Questionário

Ao contratar a coleta de dados, você deve ter vários objetivos claros em mente e fornecer orientações específicas sobre o conteúdo do instrumento de coleta de dados ou questionário. Os instrumentos de coleta de dados devem levantar todas as informações necessárias para responder à questão de política definida pela avaliação de impacto.

Desenvolvendo Indicadores

Como discutimos anteriormente, *indicadores* devem ser mensurados ao longo de toda a cadeia de resultados, incluindo indicadores de impacto final, indicadores de impacto intermediário, mensurações de entregas da intervenção, fatores exógenos e características de controle.

É importante ser seletivo sobre quais indicadores mensurar. A seletividade ajuda a limitar os custos da coleta de dados, simplifica a tarefa do órgão de coleta de dados e melhora a qualidade dos dados coletados, minimizando o tempo exigido dos entrevistados. A coleta de informação irrelevante ou de uso improvável tem um custo muito elevado. Ter um plano de análise de dados por escrito, com antecedência, ajudará a identificar as prioridades e as informações necessárias.

Os dados sobre os indicadores de resultados e características de controle devem ser coletados de forma consistente na *linha de base* e na *pesquisa de seguimento*. A coleta de dados de linha de base é altamente desejável. Mesmo com o uso de seleção aleatória ou um método de regressão descontínua,

nos quais as diferenças de pós-intervenção simples podem, em princípio, ser usadas para estimar o impacto de um programa, dados de linha de base são essenciais para testar se o desenho da avaliação de impacto é adequado (vide lista de verificação no Quadro 8.1 do Capítulo 8). Ter dados de linha de base também constitui uma 'apólice de seguro' para o caso da aleatoriedade não funcionar – nesta situação, métodos de *diferença-em-diferenças* podem ser utilizados como alternativa. Os dados de linha de base também são úteis durante o estágio da análise de impacto, uma vez que as variáveis de controle da linha de base podem ajudar a aumentar o poder estatístico e permitem analisar os impactos em diferentes subpopulações. Por fim, dados da linha de base podem ser utilizados para melhorar o desenho do programa. Por exemplo, os dados de linha de base, por vezes, tornam possível analisar a eficácia da focalização ou fornecem informações adicionais sobre os beneficiários ao órgão de execução do programa.

Medindo Indicadores

Após a definição dos principais dados que precisam ser coletados, o passo seguinte é determinar exatamente como medir esses indicadores. A *mensuração* em si é uma arte e é melhor delegá-la ao órgão contratado para a coleta de dados, aos especialistas em pesquisa ou aos avaliadores. Já foram escritos diversos livros sobre como medir, da melhor maneira possível, determinados indicadores em contextos específicos como, por exemplo, a redação exata das perguntas realizadas em pesquisas domiciliares (vide referências em Grosh e Glewwe [2000] e ONU [2005])[1] ou dos procedimentos detalhados que devem ser seguidos para coletar dados de saúde ou de testes de aprendizagem. Embora essas discussões possam parecer complicadas, são extremamente importantes. Aqui forneceremos alguns princípios gerais de orientação para guiar a contratação da coleta de dados.

Os indicadores de resultados devem ser os mais consistentes possíveis, de acordo com as melhores práticas nacionais e internacionais. É sempre útil considerar como os indicadores de interesse foram mensurados nas melhores práticas de pesquisa, tanto nacionais quanto internacionais. Utilizar os mesmos indicadores (incluindo os mesmos módulos de pesquisa ou questões) garante a comparabilidade entre dados pré-existentes e os coletados para a avaliação de impacto. Se você decidir escolher um indicador que não seja totalmente comparável ou não seja tão bem mensurado, a utilidade dos resultados da avaliação pode ser limitada.

Todos os indicadores devem ser mensurados exatamente do mesmo modo em todas as unidades, em ambos os grupos: o de tratamento e o de comparação. Utilizar métodos diferentes de coleta de dados (usar, por

exemplo, uma pesquisa por telefone para um e uma pesquisa presencial para o outro) cria o risco de gerar viés. O mesmo vale para a coleta de dados em períodos diferentes para os dois grupos (por exemplo, a coleta de dados para o grupo de tratamento durante a estação chuvosa e para o grupo de comparação durante a estação da seca). Essa é a razão pela qual os procedimentos utilizados para mensurar qualquer indicador de resultado devem ser formulados de forma muito precisa. O processo de coleta de dados deve ser exatamente o mesmo para todas as unidades. No questionário, cada módulo relacionado ao programa deve ser introduzido sem que afete o fluxo ou a estrutura de respostas em outras partes do questionário.

Formatando Questionários

Como maneiras diferentes de fazer a mesma pergunta de pesquisa podem produzir respostas diferentes, tanto a estruturação quanto o formato das questões devem ser os mesmos para todas as unidades, para evitar qualquer viés do entrevistador ou do entrevistado. Glewwe (ONU 2005) faz seis recomendações específicas sobre a formatação de questionários para pesquisas domiciliares. Estas recomendações aplicam-se, igualmente, à maioria dos outros instrumentos de coleta de dados:

1. Cada pergunta deve ser escrita na íntegra no questionário, de forma que o entrevistador possa conduzir a entrevista lendo cada pergunta palavra por palavra.
2. O questionário deve incluir definições precisas de todos os principais conceitos utilizados na pesquisa, de forma que o entrevistador possa consultar a definição durante a entrevista, se necessário.
3. Cada pergunta deve ser a mais curta e simples possível e deve usar termos comuns, do cotidiano.
4. Os questionários devem ser concebidos de forma que as respostas a quase todas as questões sejam pré-codificadas.
5. O esquema de codificação de respostas deve ser consistente para todas as perguntas.
6. A pesquisa deve incluir códigos de salto, que indicam quais perguntas não devem ser feitas em função das respostas dadas às questões anteriores.

Uma vez que um questionário tenha sido elaborado pela pessoa encarregada de trabalhar no instrumento, deve ser apresentado a uma equipe de especialistas para discussão. Todos os envolvidos na avaliação (formuladores de

políticas, pesquisadores, analistas de dados e coletores de dados) devem ser consultados sobre se o questionário coleta todas as informações desejadas de modo adequado.

Testando o Questionário

É muito importante que o questionário seja testado extensivamente de forma piloto e em campo antes de ser finalizado. *Pilotos* extensivos do questionário testarão o seu formato, bem como qualquer formatação alternativa e opções de redação das perguntas. *Testes de campo* com o questionário completo, em condições reais, são fundamentais para verificar a duração e se o formato é suficientemente consistente e abrangente para produzir medições precisas de todas as informações relevantes. Testes de campo são parte integrante do trabalho de desenho do questionário contratado.

Realizando o Trabalho de Campo

Mesmo quando você contrata a coleta de dados, a compreensão clara de todas as etapas envolvidas em tal processo é fundamental para ajudar a garantir que existam os *mecanismos de controle de qualidade* necessários e os *incentivos* corretos. A entidade responsável pela coleta de dados terá de coordenar o trabalho de um grande número de atores diferentes, incluindo os entrevistadores, supervisores, coordenadores de campo e pessoal de apoio logístico, além de uma equipe de entrada de dados, composta por programadores, supervisores e operadores de entrada de dados. Deve ser elaborado um plano de trabalho claro para coordenar o trabalho de todas essas equipes - o *plano de trabalho* é um produto chave.

No início, o plano de trabalho deve incluir *treinamento* adequado para a equipe de coleta de dados, antes da coleta começar. Deverá ser preparado um *manual de referência* completo para o treinamento, utilizado durante todo o trabalho de campo. O treinamento é fundamental para garantir que os dados sejam coletados de forma consistente por todos os envolvidos. O processo de formação também é uma boa oportunidade para identificar os entrevistadores com melhor desempenho e realizar um último teste piloto de instrumentos e procedimentos sob condições normais. Uma vez que a amostra tenha sido extraída, os instrumentos projetados e testados e as equipes devidamente capacitadas, a coleta de dados pode começar. É uma boa prática garantir que o plano de trabalho de campo faça com que cada

equipe de pesquisa colete dados no mesmo número de unidades de tratamento e de comparação.

Como discutido no capítulo 11, um processo de amostragem adequado é essencial para garantir a qualidade da amostra. No entanto, muitos erros não relacionados à amostragem podem ocorrer enquanto os dados são coletados. No contexto de uma avaliação de impacto, uma preocupação é de que os erros podem não ser os mesmos para os grupos de tratamento e de comparação.

Conceito-chave:
A não resposta surge quando faltam dados ou há dados incompletos para algumas unidades da amostra. A não resposta pode gerar viés nos resultados da avaliação.

A *não resposta* surge quando se torna impossível coletar dados completos para algumas unidades da amostra. Como as amostras reais são restritas àquelas unidades para as quais os dados podem ser coletados, unidades que optem por não responder a uma pesquisa podem tornar a amostra menos representativa e criar viés nos resultados da avaliação. *Atrição* é uma forma comum de não resposta, que ocorre quando algumas unidades abandonam a amostra entre as rodadas de coleta de dados, por exemplo, já que os migrantes não são totalmente rastreados.

A não resposta e a atrição são particularmente problemáticos no contexto das avaliações de impacto, porque podem criar diferenças entre o grupo de tratamento e o grupo de comparação. Por exemplo, a atrição pode ser diferente entre os dois grupos: se os dados estão sendo coletados depois que o programa começou a ser implementado, a taxa de resposta entre as unidades de tratamento pode ser mais elevada do que a taxa entre as unidades de comparação. Isso pode acontecer porque as unidades de comparação estão insatisfeitas por não terem sido selecionadas ou são mais propensas a migrar. Não respostas também podem ocorrer dentro do próprio questionário, normalmente porque faltam alguns indicadores ou porque os dados de determinada unidade estão incompletos.

Erro de medição é outro tipo de problema que pode gerar viés, se for sistemático. O erro de medição é a diferença entre o valor de uma característica, conforme o entrevistado e o valor verdadeiro (mas desconhecido) (Kasprzyk 2005). Essa diferença pode ser atribuída à forma como o questionário está redigido ou ao método escolhido de coleta de dados, ou pode ocorrer por causa dos entrevistadores que estão aplicando a pesquisa ou do entrevistado que está dando as respostas.

A qualidade da avaliação de impacto depende diretamente da qualidade dos dados coletados. Padrões de qualidade precisam ficar claros para todos os envolvidos no processo de coleta de dados. Os padrões deverão ser particularmente enfatizados durante o treinamento dos entrevistadores e nos manuais de referência. Por exemplo, procedimentos detalhados para minimizar a não resposta ou (se aceitável) a substituição de unidades na amostra são essenciais. O órgão de coleta de dados deve compreender claramente as

taxas aceitáveis de não resposta e de atrição. As melhores práticas de avaliação de impacto buscam manter a não resposta e a atrição abaixo de 5%. Isso pode nem sempre ser viável em populações muito móveis mas, não obstante, fornece uma base de referência útil. Os entrevistados são, por vezes, compensados para minimizar a não resposta. Qualquer que seja o caso, o contrato com o órgão de coleta de dados deve conter incentivos claros - por exemplo, maior compensação se a taxa de não resposta for inferior a 5%, ou outro limite aceitável.

Conceito-chave:
As melhores práticas de avaliação de impacto buscam manter a não resposta e a atrição abaixo de 5%.

Devem ser estabelecidos *procedimentos de garantia de qualidade* bem-definidos para todos os estágios do processo de coleta de dados, incluindo a concepção do procedimento de amostragem e do questionário, assim como os estágios de preparação, coleta, entrada, limpeza e armazenamento dos dados.

Deve ser dada alta prioridade às verificações de qualidade durante o trabalho de campo, para minimizar os erros de não resposta de cada unidade. Devem existir procedimentos claros para revisitar unidades que não forneceram informações ou forneceram informações incompletas. Vários filtros devem ser introduzidos no processo de controle de qualidade, fazendo, por exemplo, com que entrevistadores, supervisores e, se necessário, coordenadores de campo revisitem as unidades de não resposta, a fim de verificar sua situação. Os questionários de entrevistas não respondidas ainda deverão ser claramente codificados e registrados. Uma vez que os dados tenham sido completamente digitalizados, as taxas de não resposta podem ser resumidas e todas as unidades da amostra levadas em conta.

Também devem ser feitas verificações de qualidade em qualquer dado incompleto de determinada unidade pesquisada. Novamente, o processo de controle de qualidade deve incluir múltiplos filtros. O entrevistador é responsável por verificar os dados imediatamente após terem sido coletados. O supervisor e o coordenador de campo devem realizar controles aleatórios em um estágio posterior.

As verificações de qualidade para erros de medição são mais difíceis, mas são cruciais para avaliar se a informação foi coletada com precisão. Verificações de consistência podem ser incorporadas no questionário. Além disso, os supervisores precisam realizar *verificações por amostragem* e verificações cruzadas para garantir que os entrevistadores coletem dados de acordo com os padrões de qualidade estabelecidos. Coordenadores de campo também devem contribuir para que essas verificações minimizem potenciais conflitos de interesses dentro da empresa de pesquisa.

É fundamental que todas as etapas envolvidas na verificação da qualidade sejam solicitadas, explicitamente, quando da contratação da coleta de dados. Pode-se também considerar a contratação de um órgão externo para

auditar a qualidade das atividades de coleta de dados. Fazer isso pode limitar significativamente a gama de problemas que podem surgir como resultado da falta de supervisão da equipe de coleta de dados.

Processando e Validando os Dados

Normalmente, as pesquisas domiciliares são realizadas com lápis e papel, embora, mais recentemente, a coleta de dados eletrônica, usando computadores portáteis,e outros dispositivos tem se tornado mais comum. Qualquer que seja o caso, os dados devem ser digitalizados e processados. Tem de ser desenvolvido um *software de entrada de dados* e elaborado um sistema para gerenciar o fluxo de dados a ser digitalizados. Devem ser estabelecidos procedimentos e normas e os operadores de entrada de dados devem ser cuidadosamente treinados, para garantir que a entrada de dados seja consistente. Na medida do possível, a entrada de dados deve ser integrada às operações de coleta de dados (inclusive durante a fase de testes piloto), de modo que qualquer problema com os dados coletados possa ser prontamente identificado e verificado em campo.

Ao trabalhar com pesquisas com papel e lápis, a referência de qualidade para o processo de entrada de dados deve ser que os dados físicos brutos sejam replicados exatamente na versão digitalizada, sem modificações durante sua inserção na base. Para minimizar os erros de entrada de dados, é aconselhável encomendar um procedimento *duplamente cego de entrada de dados*, que pode ser utilizado para identificar e corrigir qualquer erro remanescente.

Além destas verificações da qualidade durante o processo de entrada de dados, pode ser desenvolvido um software para executar verificações automáticas de muitos erros não amostrais (de não resposta a itens e de inconsistência) que podem ocorrer no campo. Se o processo de entrada de dados for integrado aos procedimentos de trabalho de campo, podem ser remetidos dados incompletos ou inconsistentes aos trabalhadores de campo para a verificação no local (Muñoz 2005, capítulo 15). Esse tipo de integração cria desafios para o fluxo organizacional de operação de campo, mas pode render ganhos expressivos de qualidade, diminuindo erros de medição e aumentando o poder da avaliação de impacto. A possibilidade de utilizar uma abordagem integrada deve ser considerada explicitamente quando a coleta de dados estiver sendo planejada. Novas tecnologias podem facilitar o controle de qualidade.

Como vimos, a coleta de dados compreende um conjunto de operações cuja complexidade não deve ser subestimada. O Quadro 12.1 discute como o

Quadro 12.1: Coleta de Dados para a Avaliação do Piloto do Programa *Atención à Crisis* na Nicarágua

Em 2005, o governo da Nicarágua lançou o piloto do programa *Atención à Crisis*. Seu objetivo foi avaliar o impacto da combinação de um programa de transferência condicionada de renda (PTCR) com transferências produtivas, como subsídios para o investimento em atividades não agrícolas ou treinamento profissional. O piloto do *Atención a Crisis* foi implementado pelo Ministério da Família, com apoio do Banco Mundial.

Na avaliação, foi utilizada uma alocação aleatória em dois estágios. Primeiro, 106 comunidades foram aleatoriamente alocadas ao grupo de comparação ou ao grupo de tratamento. Segundo, nas comunidades de tratamento foi alocado, aleatoriamente, um dos três pacotes de benefícios aos domicílios elegíveis: (1) a transferência condicionada de renda; (2) o PCTR, mais uma bolsa de estudos que permitia que um dos membros do domicílio escolhesse entre uma série de cursos de treinamento profissional; e (3) o PCTR, mais um subsídio ao investimento produtivo, para incentivar os beneficiários a iniciar uma pequena atividade não agrícola, com o objetivo de criação de ativos e diversificação da renda (Macours e Vakis 2009).

Uma pesquisa de linha de base foi realizada em 2005, uma primeira pesquisa de seguimento em 2006 e uma segunda pesquisa de seguimento em 2008, dois anos após o fim da intervenção. Controles de qualidade rigorosos foram realizados em todos os estágios do processo de coleta de dados. Primeiro, os questionários foram cuidadosamente testados em campo e os entrevistadores foram treinados, tanto em sala de aula como em condições de campo. Segundo, foi criada uma supervisão de campo, para que todos os questionários fossem revisados várias vezes por entrevistadores, supervisores, coordenadores de campo e outros revisores. Terceiro, foi utilizado um sistema de entrada de dados duplamente cego, juntamente com um programa abrangente de verificação de qualidade que identificou questionários incompletos ou inconsistentes. Questionários com itens não respondidos ou inconsistências foram sistematicamente devolvidos ao campo para verificação. Esses procedimentos e exigências foram explicitamente especificados nos termos de referência da empresa de coleta de dados.

Além disso, foram realizados procedimentos detalhados de rastreamento, para minimizar a atrição. No início, em 2005, foi realizado um censo completo das famílias que residem nas comunidades de tratamento e de controle, em estreita colaboração com os líderes comunitários. Dada a presença de expressiva mobilidade geográfica, foram concedidos à empresa de pesquisa incentivos para acompanhar os migrantes individuais por todo o país. Como resultado, apenas 2% dos 4.359 domicílios originais não puderam ser entrevistados em 2009. A empresa de pesquisa também foi contratada para rastrear todas as pessoas dos domicílios pesquisados em 2005. Novamente, apenas 2% dos indivíduos a quem as transferências do programa foram dirigidas não puderam ser rastreados (outros 2% haviam morrido). A atrição foi de 3% para todas as crianças de domicílios pesquisados em 2005 e 5% para todos os indivíduos em domicílios pesquisados em 2005.

(continua)

Quadro 12.1 *continuação*

Taxas de atrição e de não resposta fornecem um bom indicador da qualidade da pesquisa. Chegar até essas taxas de atrição extremamente baixas exigiu intensos esforços por parte da empresa de coleta de dados, bem como incentivos explícitos. O custo unitário de uma casa ou indivíduo rastreado é também muito maior, o que precisa ser contabilizado. Além disso, os controles de qualidade rigorosos tiveram um custo e aumentaram o tempo de coleta de dados. Ainda assim, no contexto do piloto do *Atención a Crisis*, a amostra permanecia representativa, tanto em nível domiciliar quanto individual, três a quatro anos após a linha de base; o erro de medição foi minimizado e a confiabilidade da avaliação foi assegurada. Como resultado, o piloto do *Atención a Crisis* é um dos programas de rede de proteção social cuja sustentabilidade pode ser estudada de forma mais convincente.

Fonte: Macours e Vakis 2009; autores.

processo de coleta de dados para a avaliação do piloto do programa *Atención a Crisis* na Nicarágua rendeu dados de alta qualidade, com atrição e não resposta extremamente baixos e poucos erros de medição e de processamento. Dados de alta qualidade como esses somente podem ser obtidos quando procedimentos de qualidade de dados e incentivos apropriados são postos em prática no momento da contratação da coleta de dados.

No final do processo de coleta de dados, o conjunto de dados obtido deve ser entregue com documentação detalhada - incluindo uma tabela de codificação completa e um dicionário de dados - e armazenado em local seguro. Se os dados tiverem sido coletados para uma avaliação de impacto, então o conjunto de dados deve também incluir informações complementares sobre o status de tratamento e a participação no programa. Uma documentação completa acelerará a análise dos dados de avaliação de impacto, que produzirá resultados que podem ser usados para a elaboração de políticas em tempo hábil. Também facilitará o compartilhamento de informações.

Nota

1. Vide também Fink e Kosecoff (2008); Iarossi (2006) e Leeuw, Hox e Dillman (2008), que fornecem uma ampla gama de orientações práticas e importantes sobre a coleta de dados.

Referências

Fink, A. & Kosecoff, J. (2008). How to Conduct Surveys: A Step by Step Guide. 4ª ed. Londres: Sage Publications.

Glewwe, P. (2005). An Overview of Questionnaire Design for Household Surveys in Developing Countries. In *Household Sample Surveys in Developing and Transition Countries*, capítulo 3. Nova York: Nações Unidas.

Grosh, M. & Glewwe, P., eds. (2000). Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Study. Washington, DC: Banco Mundial.

Iarossi, G. (2006). The Power of Survey Design: A User's Guide for Managing Surveys, Interpreting Results, and Influencing Respondents. Washington, DC: Banco Mundial.

Kasprzyk, D. (2005). Measurement Error in Household Surveys: Sources and Measurement. In *Household Sample Surveys in Developing and Transition Countries*, capítulo 9. Nova York: Nações Unidas.

Leeuw, E., Hox J. & Dillman, D. (2008). International Handbook of SurveyMethodology. Nova York: Taylor & Francis Group.

Macours, K. e Vakis, R. (2009). Changing Household Investments and Aspirations through Social Interactions: Evidence from a Randomized Experiment. *Policy Research Working Paper 5137*, Washington, DC: Banco Mundial.

Muñoz, J. (2005). A Guide for Data Management of Household Surveys. In *Household Sample Surveys in Developing and Transition Countries,* capítulo 15. Nova York: Nações Unidas.

ONU (Nações Unidas). 2005. *Household Sample Surveys in Developing and Transition Countries.* Nova York: Nações Unidas.

CAPÍTULO 13

Produzindo e Divulgando os Resultados

Neste capítulo, discutiremos o conteúdo e o uso dos vários relatórios produzidos durante uma avaliação de impacto. Durante a fase de preparação, o administrador da avaliação, normalmente, preparará um plano de avaliação de impacto, com detalhes dos objetivos, métodos, amostras e estratégias de coleta de dados para a avaliação (o Quadro 13.1 apresenta um resumo do processo). Os vários elementos do plano de avaliação são discutidos nos Capítulos 1 a 12.

Uma vez que a avaliação esteja sendo executada, os avaliadores produzirão uma série de relatórios, incluindo o *relatório de linha de base*, o *relatório de avaliação de impacto* e *informes de políticas*. Os avaliadores também devem produzir conjuntos de dados integralmente documentados como produtos finais. Assim que o relatório da avaliação de impacto estiver disponível e os resultados forem conhecidos, será a hora de pensar na melhor maneira de divulgar as descobertas entre os formuladores de políticas e outros atores interessados no processo de desenvolvimento. A produção e a divulgação das descobertas da avaliação de impacto são o ponto principal deste capítulo.

Quadro 13.1: Estrutura de um Plano de Avaliação de Impacto

1. Introdução
2. Descrição da intervenção
3. Objetivos da avaliação
   - 3.1 Hipóteses, teoria de mudança, cadeia de resultados
   - 3.2 Questões de políticas
   - 3.3 Principais indicadores de resultado
4. Desenho da avaliação
5. Amostragem e dados
   - 5.1 Estratégia de amostragem
   - 5.2 Cálculos de poder estatístico
6. Planos de coleta de dados
   - 6.1 Pesquisa de linha de base
   - 6.2 Pesquisa(s) de acompanhamento/seguimento
7. Produtos a ser entregues
   - 7.1 Relatório de linha de base
   - 7.2 Relatório de avaliação de impacto
   - 7.3 Informes de políticas
   - 7.4 Conjunto de dados integralmente documentados
8. Plano de divulgação
9. Questões éticas
10. Linha de tempo
11. Orçamento e financiamento
12. Composição da equipe de avaliação

Que Produtos a Avaliação Entregará?

Os principais produtos de uma avaliação são um relatório de avaliação de impacto e diversos informes com análises de políticas que resumem os principais achados.

Pode levar vários anos, a partir do início da avalição, até que o relatório seja concluído, uma vez que os achados da avaliação só podem ser produzidos quando os dados de seguimento estiverem disponíveis. Devido a esta

defasagem, os formuladores de políticas geralmente solicitam produtos de avaliação intermediários, tais como o relatório de linha de base, a fim de disponibilizar informações preliminares para sustentar o diálogo e decisões sobre políticas[1].

Conforme discutido no Capítulo 10, o administrador da avaliação trabalhará com analistas de dados para produzir os relatórios de linha de base e o relatório final. Os analistas de dados são profissionais de estatística ou econometria que programarão a análise da avaliação de impacto em software estatístico, como o Stata, SPSS ou R. Os analistas de dados são responsáveis por garantir a qualidade, o rigor científico e a credibilidade dos resultados. Aqui, não discutiremos como analisar dados[2]; ao invés disso, resumiremos o escopo dos relatórios para os quais os dados contribuirão.

Produto Intermediário: Relatório de Linha de Base

Os principais objetivos de um relatório de linha de base são aferir se o desenho de avaliação de impacto escolhido será válido na prática e descrever as características de linha de base (pré-programa) e resultados da população elegível. Um relatório de linha de base também gera informações sobre o programa e sobre os seus beneficiários, que podem ser úteis para melhorar tanto a execução do programa quanto a sua avaliação. O Quadro 13.2 resume o conteúdo sugerido de um relatório de linha de base[3].

O relatório de linha de base é produzido a partir da análise do conjunto tratado de dados de linha de base, complementado por dados administrativos sobre o estado de tratamento de cada unidade de análise. A designação de famílias, indivíduos ou instalações para o tratamento ou para o grupo de comparação é, geralmente, realizada após a coleta dos dados de linha de base. Desta forma, o estado de tratamento é, geralmente, registrado em um conjunto separado de dados administrativos. Pode ser organizado um sorteio público, por exemplo, para determinar quais comunidades, entre todas as elegíveis onde a pesquisa de linha de base foi realizada, serão beneficiadas com um programa de transferência de renda. Para isso ser feito, os analistas de dados devem unir os dados administrativos aos dados da linha de base. Se a avaliação incluir mais do que, digamos, 100 unidades elegíveis, não será prático fundir os dados de linha de base com os dados administrativos por nome. A cada unidade elegível precisará ser atribuído um número exclusivo, ou identificador, que a identificará em todas as fontes de dados, incluindo as bases de dados administrativos e de linha de base.

As primeiras seções do relatório de linha de base se baseiam no plano de avaliação de impacto, apresentando a motivação para a avaliação, a

Quadro 13.2: Estrutura de um Relatório de Linha de Base

1. Introdução
2. Descrição da intervenção (benefícios, regras de elegibilidade e assim por diante)
3. Objetivos da avaliação
   - 3.1 Hipóteses, teoria de mudança, cadeia de resultados
   - 3.2 Questões de políticas
   - 3.3 Principais indicadores de resultado
4. Desenho da avaliação
   - 4.1 Desenho original
   - 4.2 Participantes de fato do programa e não participantes
5. Amostragem e dados
   - 5.1 Estratégia de amostragem
   - 5.2 Cálculos de poder estatístico
   - 5.3 Dados coletados
6. Validação do desenho de avaliação
7. Estatísticas descritivas abrangentes
8. Conclusão e recomendações de implementação

descrição da intervenção (incluindo os benefícios e as regras de atribuição dos benefícios), os objetivos da avaliação (incluindo a teoria de mudança, questões básicas de política, hipóteses e indicadores) e o desenho da avaliação. A seção do desenho da avaliação deve discutir se a atribuição dos benefícios do programa foi realizada de forma consistente com o método planejado. Por ser a atribuição geralmente realizada após a conclusão da pesquisa de linha de base, é uma boa prática incluir no relatório de linha de base informações relativas à atribuição efetiva. A seção sobre amostragem geralmente começa com o resumo da estratégia de amostragem e os cálculos de poder estatístico gerados para o plano de avaliação, antes de descrever, detalhadamente, como os dados de linha de base foram coletados e o tipo de informação disponível. O relatório deve discutir os desafios enfrentados durante a coleta de dados de linha de base e deve apresentar os principais indicadores de qualidade dos dados, como as taxas de não resposta. Em relação a isso, o relatório de linha de base destacará questões que devem ser

abordadas no seguimento. Por exemplo, se a taxa de não resposta for alta na linha de base, os avaliadores precisarão desenvolver novos procedimentos para o rastreamento ou trabalho de campo, para garantir que isto não aconteça novamente durante a pesquisa de seguimento.

Como já dissemos, o principal objetivo do relatório de linha de base é fornecer uma avaliação antecipada da validade do desenho da avaliação, na prática. No Capítulo 8, destacamos que a maioria dos métodos de avaliação de impacto produzem estimativas válidas do contrafactual somente sob pressupostos específicos. O Quadro 8.1 (Capítulo 8) apresenta uma lista de testes que podem ser usados para avaliar se determinado método é apropriado em um determinado contexto. Alguns destes não requerem dados de seguimento e podem ser aplicados assim que os dados de linha de base estiverem disponíveis. Por exemplo, se for usada a alocação aleatória ou o método de oferta aleatória, o relatório de linha de base deve determinar se os grupos de tratamento e de comparação apresentam características semelhantes de linha de base. Se a avaliação se baseia no método de regressão descontínua, o relatório de linha de base deve reportar testes da continuidade do índice de elegibilidade em torno do ponto de corte. Embora estes testes de falseamento não garantam que o grupo de comparação permanecerá valido até a pesquisa de seguimento, é fundamental que o relatório de linha de base os documente.

Além de testar a validade do desenho de avaliação, o relatório de linha de base deve incluir tabelas que descrevam as características da amostra de avaliação. Elas podem melhorar a implementação do programa, ao permitir que os administradores do programa entendam melhor o perfil dos beneficiários e adaptem a intervenção do programa às suas necessidades. Por exemplo, ao conhecer o nível de educação ou experiência média de trabalho dos participantes em um programa de treinamento, os administradores do programa podem ajustar o conteúdo dos cursos de treinamento.

Do ponto de vista da avaliação, a pesquisa de linha de base geralmente produz informações que não estavam disponíveis no momento em que o plano de avaliação estava sendo elaborado. Digamos que você esteja avaliando o impacto de um programa comunitário de saúde na diarreia infantil. Ao redigir o plano de avaliação, você pode não conhecer a incidência da diarreia nesta comunidade. Então, no plano de avaliação, você teria apenas uma estimativa e basearia seus cálculos de poder estatístico nesta estimativa. No entanto, uma vez que você tenha os dados de linha de base, você poderá verificar a incidência real de diarreia na linha de base da diarreia e, consequentemente, se o tamanho da sua amostra original é adequado. Se você descobrir que os valores de linha de base dos

indicadores de resultados são diferentes daqueles usados para executar os cálculos originais de poder estatístico, o relatório de linha de base deve, então, incluir cálculos com valores atualizados.

Para assegurar a credibilidade dos resultados finais da avaliação, é boa prática permitir que profissionais externos avaliem o relatório de linha de base. Divulgar o relatório de linha de base também pode reforçar o diálogo político entre as partes interessadas em todo o ciclo da avaliação.

Produtos Finais: Relatório da Avaliação de Impacto, Informes de Políticas e Conjunto de Dados

O relatório final da avaliação de impacto é o principal produto de uma avaliação e é produzido após a coleta dos dados de seguimento[4]. Os principais objetivos do relatório de avaliação são apresentar os resultados da avaliação e responder a todas as questões de políticas feitas inicialmente. Como complemento, o relatório também precisa mostrar que a avaliação se baseia em estimativas válidas do contrafactual e que os impactos estimados são completamente atribuíveis ao programa.

O relatório final da avaliação de impacto é um relatório abrangente, que resume todo do trabalho associado à avaliação e inclui descrições detalhadas da análise de dados e especificações econométricas, assim como uma discussão de resultados, tabelas e apêndices. O Quadro 13.3 resume o conteúdo de um relatório completo de uma avaliação de impacto. Muitos bons exemplos de relatórios finais de avaliação de impacto estão disponíveis: Maluccio e Flores (2005), Levy e Ohls (2007) ou Skoufias (2005), sobre programas de transferência condicionada de renda; Card et al. (2007), sobre programas de capacitação de jovens; Cattaneo et al. (2009), sobre programas de habitação; e Basinga et al. (2010), sobre programas de financiamento com base em resultados para o setor da saúde.

A partir do relatório de linha de base, os avaliadores trabalharão com os analistas de dados para produzir o relatório final da avaliação de impacto. Os analistas começarão produzindo uma base de de dados mestre, contendo os dados de linha de base, os dados de seguimento, os dados administrativos de implementação do programa e os dados da alocação original para os grupos de tratamento e de comparação. Todas essas fontes devem ser reunidas usando-se um identificador exclusivo para cada unidade de análise.

Visto que o relatório final da avaliação de impacto é o principal produto da avaliação, ele deve incorporar as principais informações apresentadas no plano de avaliação e no relatório de linha de base, antes de tratar da análise e discussão dos resultados. Em particular, a parte introdutória do relatório final deve apresentar a lógica completa da intervenção e da avaliação,

Quadro 13.3: Estrutura de um Relatório de Avaliação

1. Introdução
2. Descrição da intervenção (benefícios, regras de elegibilidade e assim por diante).
   - 2.1 Desenho
   - 2.2 Implementação
3. Objetivos da avaliação
   - 3.1 Hipóteses, teoria de mudança, cadeia de resultados
   - 3.2 Questões de políticas
   - 3.3 Principais indicadores de resultado
4. Desenho da avaliação
   - 4.1 Na teoria
   - 4.2 Na prática
5. Amostragem e dados
   - 5.1 Estratégia de amostragem
   - 5.2 Cálculos de poder estatístico
   - 5.3 Dados coletados
6. Validação do desenho de avaliação
7. Resultados
8. Testes de robustez
9. Conclusão e recomendações de políticas

bem como descrever a intervenção (benefícios e regras de alocação dos benefícios), os objetivos da avaliação (incluindo a teoria de mudança, as questões básicas de política, hipóteses e indicadores), o desenho original da avaliação e como ela foi implementada na prática.

Em geral, a interpretação dos resultados depende, fundamentalmente, de como a intervenção foi implementada. O relatório final da avaliação deve, portanto, discutir detalhadamente a implementação da intervenção. Isto pode ser feito antes da apresentação dos resultados, por meio da descrição dos dados sobre a implementação do programa, obtidos de pesquisas de seguimento ou fontes administrativas complementares.

A seção de amostragem e dados é o lugar onde a estratégia de amostragem e os cálculos de poder estatístico são descritos, antes da discussão extensiva da coleta dos dados de linha de base e de seguimento.

Os principais indicadores de qualidade dos dados, tais como as taxas de não resposta e de atrição, devem ser apresentados para cada rodada de dados. Se as taxas de não resposta e de atriçãoforem altas, é fundamental que o analista de dados discuta como isso pode afetar a interpretação dos resultados. Por exemplo, é fundamental testar se a atriçãoe a não resposta estão equilibradas entre os grupos de comparação e o tratamento.

Uma vez que os dados tenham sido descritos, o relatório pode tratar da apresentação dos resultados referentes a todas as principais questões de políticas e os indicadores de resultado identificados como objetivos da avaliação. A estrutura da seção de resultados dependerá dos tipos de questões de políticas estudadas. Por exemplo, a avaliação testa várias alternativas de programa, ou testa somente se uma intervenção funciona ou não? Os formuladores de políticas pediram uma análise de como os resultados variam entre os subgrupos? Para as avaliações que foram bem elaboradas e implementadas, resultados rigorosos de avaliação podem, geralmente, ser apresentados de maneira intuitiva.

Como já dissemos, o relatório de avaliação de impacto deve fornecer forte evidência de que os impactos estimados são integralmente atribuídos ao programa. Portanto, o relatório deve examinar com cuidado a validade do desenho da avaliação. Para demonstrar a validade do desenho da avaliação de impacto, o primeiro passo é apresentar os resultados dos testes de falseamento executados com dados de linha de base (Quadro 8.1, Capítulo 8). O relatório também deve conter os resultados de qualquer teste que possa ser executado com os dados de seguimento. Por exemplo, se uma abordagem de *diferença-em-diferenças* for escolhida, a série de testes de falseabilidade descrita no Quadro 8.1 pode ser executada somente na presença de dados de seguimento.

A seção introdutória do relatório de avaliação deve documentar quaisquer novos desafios com o método de avaliação que tenham surgido entre as pesquisas de linha de base e de seguimento. Por exemplo, a não-conformidade com as regras de alocação entre grupos de tratamento e de comparação tem implicações importantes para a análise e interpretação dos resultados e deve ser discutida na parte inicial do relatório. O relatório também deve apresentar informações sobre quantas unidades designadas ao grupo de tratamento de fato receberam o programa e quantas daquelas designadas ao grupo de comparação não o receberam. Se ocorrer qualquer desvio da alocação original do programa, a análise tem que ser ajustada para compensar a não conformidade (vide as técnicas discutidas na parte 2).

Em paralelo aos testes da validação do desenho de avaliação, o relatório final deve fornecer uma discussão abrangente sobre a natureza, confiabilidade e robustez dos resultados. Ele deve conter uma série de testes de

robustez relevantes para a metodologia de avaliação sendo utilizada. Quando os métodos de pareamento são aplicados, por exemplo, o relatório deve conter resultados a partir da aplicação de técnicas alternativas, visando identificar o melhor pareamento para cada observação tratada. O analista de dados é responsável pela identificação e apresentação dos testes de robustez mais apropriados para avaliações específicas. As partes finais do relatório devem responder claramente a cada questão de política que a avaliação se propôs a responder e fornecer recomendações detalhadas de políticas com base nos resultados.

Entender como a intervenção foi realizada é particularmente fundamental quando os resultados da avaliação apresentam um impacto limitado ou negativo. Os resultados negativos - ou a ausência dos mesmos - não justificam punir os gestores do programa ou da avaliação. Ao contrário, eles oferecem uma oportunidade aos gestores do programa e da avaliação de explicarem, claramente, o que não funcionou como planejado; isto, por si só, pode trazer grandes ganhos para a política e deve ser recompensado. A comunicação contínua entre a equipe de avaliação e os formuladores de política responsáveis pelo programa é particularmente crítica quando há sinais de que uma avaliação produzirá resultados negativos ou nulos. Avaliações de processo complementares ou um trabalho qualitativo podem fornecer uma explicação valiosa sobre o motivo de um programa não ter atingido os resultados desejados. A ausência de resultados que possa ser vinculada a situações de execução imperfeita do programa deve ser claramente distinguida da falta de resultados de um programa bem executado, mas que teve um desenho fraco[5]. Em geral, as avaliações que testam alternativas do programa são mais úteis para esclarecer qual desenho de programa funciona e qual não funciona.

De uma maneira geral, a análise final dos dados deve fornecer evidências convincentes de que os impactos estimados do programa são, realmente, causados pela intervenção. Para garantir que os resultados sejam completamente objetivos e assegurar, assim, a sua legitimidade, todos os relatórios devem ser revisados por especialistas e passar por amplas consultas antes de serem finalizados. O conteúdo do relatório final da avaliação de impacto pode, subsequentemente, ser transformado em artigos acadêmicos mais técnicos, para publicação em periódicos revisados por pares, agregando credibilidade adicional aos resultados da avaliação.

Além do relatório completo de avaliação, os avaliadores devem produzir um ou mais informes curtos, com análises de questões de política para ajudar a comunicar os resultados a gestores e outras partes interessadas. Um informe de política apresenta os principais achados da avaliação por meio de gráficos, mapas e outros formatos acessíveis, bem como uma

discussão de recomendações de políticas. Também contém um breve resumo dos aspectos técnicos da avaliação. O informe de política pode ser disponibilizado para o público em forma impressa ou pela internet, circulando entre os políticos, sociedade civil e a imprensa. Bons exemplos de informes de políticas podem ser encontrados no *The Poverty Action Lab* (J-PAL) ou no site de Desenvolvimento Humano do Banco Mundial (por exemplo, Poverty Action Lab 2008; Desenvolvimento Humano do Banco Mundial 2010).

O último produto importante de uma avaliação de impacto é um conjunto de dados relevantes e sua documentação. Ferramentas como o *Microdata Management Toolkit,* da International Household Survey Network (http://www.ihsn.org), podem auxiliar neste processo. Os formuladores de política e avaliadores de impacto normalmente chegam a um acordo quanto a um cronograma no qual a análise de impacto inicial é conduzida e os dados de avaliação são liberados ao domínio público. Disponibilizar os dados publicamente aumenta a transparência, pois os resultados podem ser replicados e validados externamente. O acesso público também motivará aos pesquisadores externos a realizar análises adicionais com os mesmos dados, o que pode fornecer informações e aprendizagens valiosas para o programa. Ao disponibilizar os dados publicamente, é importante garantir o anonimato a todos os sujeitos da pesquisa; qualquer informação que possa identificar os entrevistados (como nomes, endereços ou localização) deve ser removida dos dados publicamente disponíveis. Esse tipo de informação sensível deve ser mantida em segurança e disponibilizada somente às atividades futuras de coleta autorizada de dados.

Como Divulgar os Achados?

Além da entrega dos resultados de avaliação, o último objetivo das avaliações de impacto é tornar as políticas públicas mais efetivas e melhorar os resultados de desenvolvimento. Para garantir que a avaliação de impacto informe as decisões sobre políticas, ela deve se comunicar de forma clara com todas as partes interessadas, incluindo os formuladores de política, a sociedade civil e a imprensa. Avaliações influentes geralmente incluem um plano de divulgação detalhado, que descreve como os principais atores serão informados e envolvidos durante todo o ciclo da avaliação. Tal plano de divulgação pode facilitar o uso dos resultados na formulação de políticas e garantir que as avaliações de impacto realmente atinjam os resultados.

Nas etapas iniciais do desenho de uma avaliação, os avaliadores têm a sua primeira oportunidade de construir canais fortes de comunicação com os gestores de programas.

Como deve ter ficado claro da nossa discussão sobre os métodos de avaliação, um desenho de avaliação depende diretamente de como o programa em si é elaborado e implementado e, assim, é fundamental que os avaliadores externos e gestores que contratam a avaliação colaborem durante a fase de desenho do programa. Uma equipe de avaliação que trabalhe bem garantirá que a avaliação esteja integralmente alinhada às necessidades dos gestores e que seu progresso e resultados lhes sejam comunicados regularmente.

O plano de divulgação deve detalhar como a equipe de avaliação aumentará a demanda por resultados da avaliação e como maximizará seu uso na tomada de decisões. No mínimo, os avaliadores devem fomentar a sensibilização sobre a avaliação, comunicando efetivamente os resultados às partes interessadas internas e externas em todo o ciclo da avaliação. No início da avaliação, um pré-estudo e oficina de lançamento com implementadores e atores fundamentais podem ajudar a formar consenso sobre os principais objetivos, questões de política e aspectos de desenho. Além de fornecer uma plataforma para consultas e assegurar que a avaliação esteja integralmente alinhada às necessidades das partes interessadas, tal evento é importante para conscientizar sobre a avaliação e reforçar o interesse em conhecer os resultados.

Durante a avaliação, reuniões periódicas de um comitê interinstitucional ou mesa redonda permanente de discussão podem ajudar a garantir que o trabalho da equipe de avaliação permaneça relevante para a política. Tais fóruns de discussão podem fornecer feedback e orientações para a produção de termos de referência, conteúdo do instrumento de pesquisa, divulgação dos resultados, ou sobre os canais mais apropriados para se atingir os tomadores de decisão de alto escalão.

É importante a organização de eventos de divulgação sobre produtos intermediários, tais como o relatório de linha de base, para manter um diálogo de política ativo com os usuários da avaliação. Incentivar discussões antecipadas sobre o relatório de linha de base é benéfico, tanto para divulgar resultados intermediários relevantes para a política em questão quanto para garantir a conscientização contínua sobre a natureza dos resultados futuros da avaliação de impacto.

Antes de finalizar o relatório de avaliação, alguns avaliadores optam por organizar um evento final de consulta, para dar às partes interessadas a oportunidade de comentar os resultados. Estas consultas podem contribuir para a melhoria da qualidade dos resultados da avaliação, assim como

sua aceitação. Uma vez que o relatório final da avaliação de impacto e os informes de política associados estiverem disponíveis, eventos de divulgação de alta visibilidade são fundamentais para garantir a ampla conscientização dos resultados dentre as partes interessadas. Uma consulta no país e uma oficina de divulgação com um conjunto abrangente de atores oferecem uma plataforma para discutir os resultados, reunir *feedback* e detalhar as mudanças de política que poderiam ser feitas como resultado da avaliação. Esta oficina pode ser seguida de um *workshop* de divulgação de alto nível, envolvendo os principais formuladores de política (vide quadro 13.4).

No exterior, os resultados podem ser divulgados em conferências, seminários e outras reuniões, se os resultados da avaliação forem úteis para a formulação de políticas em outros países. Outros canais inovadores de disseminação, como as interfaces da internet, também são úteis para aumentar a visibilidade dos achados.

De maneira geral, a disseminação dos produtos da avaliação de impacto, conforme um plano bem concebido e abrangendo todo o ciclo da avaliação, é importante para garantir que os resultados alimentem efetivamente o diálogo sobre a política. Somente quando os resultados da avaliação forem adequadamente compartilhados com os formuladores de políticas e os gestores de programas e usados integralmente no processo de tomada de decisão, as avaliações de impacto poderão atingir seu objetivo final, de melhorar a eficácia dos programas sociais.

Quadro 13.4: Disseminando os Achados da Avaliação para Melhorar a Política

A avaliação do financiamento com base em resultados de saúde em Ruanda fornece um bom exemplo de estratégia de divulgação bem sucedida. Sob a liderança do Ministério da Saúde, foi formada uma equipe composta de acadêmicos locais e profissionais do Banco Mundial para liderar a avaliação. Vários atores foram envolvidos em toda a avaliação, desde o lançamento, o que provou ser chave para garantir o seusucesso e um forte apoio político. Os resultados finais da avaliação (Basinga et al. 2010) foram revelados durante um evento de divulgação pública de um dia inteiro envolvendo tomadores de decisão de alto escalão e várias partes interessadas. Graças a estes canais de comunicação, os achados influenciaram fortemente a elaboração da política de saúde em Ruanda. Os resultados também foram divulgados em conferências internacionais sobre a saúde e através de um site na internet.

Fonte: Morgan 2010.

Notas

1. Uma avaliação pode gerar outros produtos intermediários. Por exemplo, trabalho qualitativo em campo ou avaliações de processo fornecem informações complementares altamente valiosas antes que o relatório final da avaliação de impacto seja produzido. O nosso foco foi o relatório de linha de base porque ele constitui o principal produto intermediário das avaliações quantitativas de impacto, o assunto deste livro.
2. Khandker et al. (2009) apresentam uma introdução à avaliação que inclui uma revisão de análise de dados e os comandos Stata relevantes para cada método de avaliação de impacto.
3. O resumo é indicativo e pode ser adaptado dependendo da natureza de cada avaliação, por exemplo, ao modificar a ordem ou conteúdo das várias seções.
4. Em casos com rodadas múltiplas de coleta de dados de seguimento, pode ser produzido um relatório de avaliação de impacto para cada rodada e os resultados podem ser comparados, para destacar se os impactos do programa são sustentáveis ou variam com duração de exposição.
5. Conforme discutido no Capítulo 1, esta é uma razão pela qual os testes de eficácia para minimizar os desafios da implementação são úteis para determinar se um desenho específico de programa funciona em circunstâncias ideais. Uma vez que se documente a prova de conceito, o piloto poderá ser expandido.

Referências

Basinga, P., Gertler, P., Binagwaho, A., Soucat, A., Sturdy, J. & Vermeersch, C. (2010). Paying Primary Health Care Centers for Performance in Rwanda. *Policy Research Working Paper Series 5190*. Washington, DC: Banco Mundial.

Card, D., Ibarraran, P., Regalia, F., Rosas, D. & Soares, Y. (2007). The Labor Market Impacts of Youth Training in the Dominican Republic: Evidence from a Randomized Evaluation. *NBER Working Paper 12883*. Washington, DC: National Bureau of Economic Research.

Cattaneo, M., Galiani, S., Gertler, P., Martinez, S. & Titiunik, R. (2009). Housing, Health and Happiness. *American Economic Journal: Economic Policy* 1 (1): 75–105.

Khandker, S., Koolwal, G. & Samad, H. (2009). Handbook on Impact Evaluation: Quantitative Methods and Practices. Washington, DC: Banco Mundial.

Levy, D. & Ohls, J. (2007). Evaluation of Jamaica's PATH Program: Final Report. Ref. Nº 8966-090. Washington, DC: Mathematica Policy Research Inc.

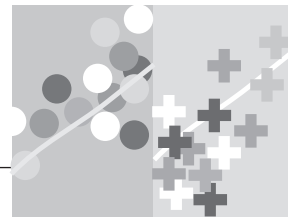
Maluccio, J. & Flores, R. (2005). Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Proteccion Social. Relatório de pesquisa 141. Washington, DC: International Food Policy Research Institute.

Morgan, L. (2010). Signed, Sealed, Delivered? Evidence from Rwanda on the Impact of Results-Based Financing for Health. *HRBF Policy Review*. Washington, DC: Banco Mundial.

Poverty Action Lab. (2008). Solving Absenteeism, Raising Test Scores. Policy Briefcase 6. Disponível em: http://www.povertyactionlab.org.

Skoufias, E. (2005). PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico. Relatório de pesquisa 139. Washington, DC: International Food Policy Research Institute.

Rede de Desenvolvimento Humano do Banco Mundial. (2010). Does Linking Teacher Pay to Student Performance Improve Results? Policy Note Series 1. Washington DC: Banco Mundial. Disponível em: http://www.worldbank.org/hdchiefeconomist.



CAPÍTULO 14

Conclusão

Este livro é um guia prático para elaborar e implementar avaliações de impacto. Esperamos que seu conteúdo desperte o interesse de três públicos principais: (1) os formuladores de políticas que consomem as informações geradas pelas avaliações de impacto; (2) gerentes de projeto e profissionais de desenvolvimento que contratam avaliações; e (3) técnicos que elaboram e implementam avaliações de impacto. Essencialmente, a avaliação de impacto diz respeito à geração de evidências sobre quais políticas sociais funcionam e quais não funcionam. Isso pode ser feito em uma estrutura clássica de avaliação de impacto, comparando resultados com e sem o programa. As avaliações de impacto também podem ser realizadas para explorar alternativas de implementação de um programa ou considerar diferentes programas, a fim de avaliar seus desempenhos comparativamente.

Discutimos que as avaliações de impacto são um investimento válido para muitos programas e que, quando associadas ao monitoramento e a outras formas de avaliação, elas permitem um entendimento claro da eficácia de políticas sociais específicas. Apresentamos um menu de metodologias de avaliação de impacto, cada uma com seu próprio conjunto de custos e benefícios com respeito à implementação, economia política, requisitos financeiros e interpretação de resultados. Discutimos que deve ser escolhido o melhor método para atender ao contexto operacional, não o contrário. Finalmente, fornecemos dicas práticas, ferramentas e orientação para

auxiliar durante o processo de avaliação e tirar o máximo proveito dos resultados da avaliação.

As avaliações de impacto são empreendimentos complexos, com muitas peças móveis. A seguinte lista destaca os principais elementos de uma avaliação de impacto bem elaborada, que deve incluir o seguinte:

- ✓ Uma questão de política concreta - fundamentada em uma teoria de mudança - que pode ser respondida com uma avaliação de impacto.
- ✓ Uma estratégia válida de identificação, que seja consistente com as regras operacionais do programa e que mostre a relação causal entre o programa e os resultados de interesse.
- ✓ Uma amostra com poder estatístico, capaz de detectar impactos relevantes da política e representativa o suficiente para permitir que os resultados sejam generalizados para a população de interesse.
- ✓ Uma fonte de dados de alta qualidade, que forneça as variáveis apropriadas requeridas pela análise, referentes aos grupos de tratamento e de comparação, usando dados de linha de base e de seguimento.
- ✓ Uma equipe de avaliação bem formada, que funcione próxima aos formuladores de política e à equipe do programa.
- ✓ Um relatório de impacto e informes associados a políticas, a ser disseminados tempestivamente às partes interessadas, alimentando tanto o desenho do programa quanto os diálogos sobre as políticas.

Também destacamos algumas dicas fundamentais que ajudam a mitigar riscos comuns inerentes ao processo de realização de uma avaliação de impacto:

- ✓ As avaliações de impacto são mais bem elaboradas se antecipadas no ciclo do projeto, idealmente como parte do desenho do programa. O planejamento antecipado permite um desenho de avaliação prospectiva baseado na melhor metodologia disponível. Também provê o tempo necessário para que se planeje e execute a coleta de dados de linha de base nas áreas de avaliação, antes do início do programa.
- ✓ Os resultados de impacto devem ser informados pela avaliação de processo e por dados rigorosos de monitoramento que passam uma imagem clara da implementação do programa. Quando os programas são bem sucedidos, é importante entender o porquê. Quando os programas falham, é importante distinguir entre um programa mal implementado e um desenho de programa deficiente.

- ✓ Coletar dados de linha de base e construir uma metodologia alternativa na elaboração de sua avaliação de impacto. Se o desenho original da avaliação for invalidado - por exemplo, caso o grupo de comparação original receba os benefícios do programa - ter um plano alternativo pode ajudar a evitar que a avaliação inteira tenha que ser descartada.
- ✓ Manter identificadores comuns entre fontes de dados diferentes, a fim de que possam ser vinculados facilmente durante a análise. Por exemplo, uma mesma família deve ter o mesmo identificador nos sistemas de monitoramento e nas pesquisas de linha de base e de seguimento.
- ✓ As avaliações de impacto são úteis tanto para conhecer o funcionamento de programas e testar alternativas programáticas quanto para avaliar o impacto geral de um conjunto de bens e serviços. Ao desagregar um programa, mesmo os grandes programas universais têm muito a aprender com a teste de inovações através de avaliações de impacto bem elaboradas. Inserir uma inovação de programa como um pequeno piloto no contexto de uma avaliação maior pode alavancar a avaliação e produzir informações valiosas para a tomada futura de decisões.
- ✓ As avaliações de impacto devem ser consideradas como mais um componente da operação do programa e devem ser orçadas e equipadas adequadamente com os recursos técnicos e financeiros necessários. Seja realista quanto aos custos e complexidade de executar uma avaliação de impacto. O processo de elaboração de uma avaliação e a coleta inicial de dados de linha de base levará, em geral, um ano ou mais. Uma vez que o programa tenha início, a intervenção precisará de um período de exposição longo o suficiente para afetar os resultados. Dependendo do programa, isso pode levar de um a cinco anos, ou até mais. A realização de uma ou mais pesquisas de seguimento, da análise e da divulgação requerem um grande esforço, de mais alguns meses. Do início ao fim, um ciclo completo de avaliação de impacto leva, geralmente, pelo menos três a quatro anos de intenso trabalho e envolvimento. São necessários recursos financeiros e técnicos adequados a cada etapa do caminho.

No final das contas, as avaliações de impacto fornecem respostas concretas a perguntas específicas de políticas. Embora essas respostas forneçam informações adaptadas para o organismo específico que contrata e paga pela avaliação, elas também fornecem informações valiosas para outros formuladores de política ao redor do mundo, que podem aprender e tomar decisões baseadas em tais evidências. Por exemplo, os mais recentes programas de transferência condicionada de renda na África e na Europa aprenderam lições das avaliações originais dos programas *Familias en Acción*, da Colômbia,

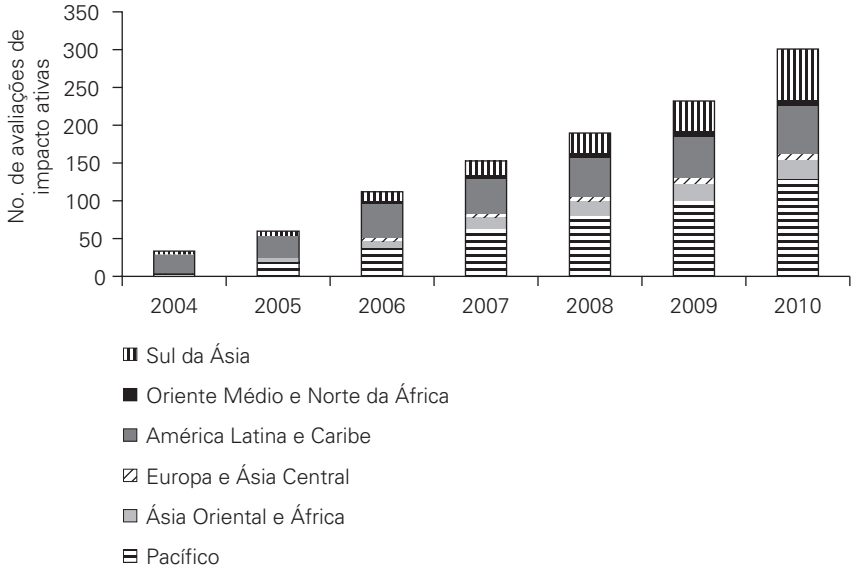
Progresa, do México e outros programas de transferência condicionada de renda na América Latina implementados ao longo dos últimos anos. Dessa forma, as avaliações de impacto são, em parte, um bem público global. Evidências geradas através de uma avaliação de impacto agregam ao conhecimento global sobre o tema. A base desse conhecimento pode, então, informar decisões políticas em outros países e contextos. Na verdade, a comunidade internacional está se posicionando para expandir seu apoio a avaliações rigorosas.

Em nível nacional, governos mais sofisticados e exigentes estão procurando demonstrar resultados e ser mais responsáveis com seus eleitores. Cada vez mais, avaliações vêm sendo realizadas por ministérios de governos nacionais e subnacionais, bem como por agências governamentais estabelecidas para liderar a agenda nacional de avaliação, tais como o Conselho Nacional para Avaliação de Políticas de Desenvolvimento Social (CONEVAL), no México, e o Departamento de Execução de Monitoramento e Avaliação, na África do Sul. Cada vez mais, evidências oriundas de avaliações de impacto estão informando as alocações orçamentárias feitas por parlamentares em nível nacional. Em sistemas onde os programas são julgados com base em evidências e resultados finais, os programas com uma forte base de evidências poderão continuar seu desenvolvimento, enquanto programas que não contam com tais provas de sucesso terão mais dificuldades em sustentar seu financiamento.

As instituições multilaterais, como o Banco Mundial e os bancos de desenvolvimento regional, assim como os organismos de desenvolvimento nacional, governos doadores e instituições filantrópicas também estão exigindo mais e melhores evidências no uso efetivo dos recursos de desenvolvimento. Estas evidências são necessárias para prestar contas àqueles que emprestam ou doam recursos, assim como para tomar decisões sobre onde é melhor alocar os recursos escassos disponíveis para o desenvolvimento. O número de avaliações de impacto realizadas por instituições de desenvolvimento aumentou consideravelmente nos últimos anos. Para ilustrar isso, a Figura 14.1 representa a quantidade de avaliações de impacto ativas ou concluídas pelo Banco Mundial entre os anos de 2004 e 2010, por região. Esta tendência positiva provavelmente continuará.

Vem surgindo um número crescente de instituições dedicadas primeiramente à produção de avaliações de impacto de alta qualidade, inclusive na arena acadêmica, como o *Poverty Action Lab*, *Innovations for Poverty Action* e o *Center of Evaluation for Global Action*, bem como organismos independentes que apoiam avaliações de impacto, tais como a *International Initiative for Impact Evaluation*. Uma série de associações relacionadas à avaliação de impacto reúnem, agora, grupos de profissionais de avaliação, pesquisadores

Figura 14.1 Número de Avaliações de Impacto no Banco Mundial por Região, 2004–10



Fonte: Banco Mundial.

e gestores de políticas interessados no assunto, incluindo o *Network of Networks on Impact Evaluation* e associações regionais, como a *African Evaluation Association* e a *Latin American and Caribbean Economics Association Impact Evaluation Network*. Todos esses esforços refletem a importância crescente da avaliação de impacto na política de desenvolvimento internacional.[1]

Dado esse crescimento nas avaliações de impacto, independentemente de se realizar avaliações profissionalmente, contratar avaliações ou usar os resultados das avaliações de impacto para a tomada de decisões, familiarizar-se com a linguagem das avaliações de impacto é cada vez mais indispensável para qualquer profissional do desenvolvimento. Evidência rigorosa, do tipo gerado através de avaliações de impacto, pode ser um dos fios condutores do diálogo da política de desenvolvimento, fornecendo subsídios para o apoio ou oposição a investimentos em programas e políticas de desenvolvimento. As evidências das avaliações de impacto permitem aos gestores de programas tomar decisões informadas sobre como conseguir resultados mais custo-efetivos. Armado com evidências de uma avaliação de impacto, o formulador de políticas tem o trabalho de fechar o ciclo, alimentando o processo de tomada de decisão com estes resultados. Esse tipo de

evidência pode informar debates, opiniões e, no final das contas, as decisões quanto à alocação de recursos humanos e financeiros pelos governos, instituições multilaterais e doadores.

Formular políticas com base em evidências significa, fundamentalmente, reprogramar orçamentos para expandir programas custo-efetivos, cortar programas ineficientes e introduzir melhorias no desenho de programas, com base nas melhores evidências disponíveis. A avaliação de impacto não é um empreendimento puramente acadêmico. As avaliações de impacto são orientadas pela necessidade de respostas a questões de políticas que afetam a vida diária das pessoas. Decisões sobre como gastar recursos escassos em programas contra a pobreza, de saúde, educação, proteção social, microcrédito, agricultura e uma miríade de outras iniciativas de desenvolvimento têm o potencial de melhorar o bem-estar da população no mundo inteiro. É vital que essas decisões sejam tomadas com o uso de evidências mais rigorosas possíveis.

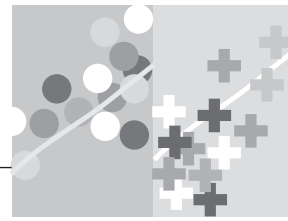
Nota

1. Para leitura adicional, vide Savedoff, Levine e Birdsall (2006).

Referências

Legovini, A. (2010). Development Impact Evaluation Initiative: A World Bank–Wide Strategic Approach to Enhance Development Effectiveness. *Draft Report for Operational Vice Presidents*. Washington, DC: Banco Mundial.

Savedoff, W., Levine, R. & Birdsall, N. (2006). When Will We Ever Learn? Improving Lives through Impact Evaluation. CGD Evaluation Gap *Working Group Paper, Center for Global Development*. Washington, DC. Disponível em: http://www.cgdev.org/content/publications/detail/7973.



GLOSSÁRIO

Itálico indica termos que são definidos no glossário.

Alocação aleatória ou desenho de controle aleatório. A alocação aleatória é considerada o método mais robusto para estimar *contrafatuais* e é frequentemente referida como o "padrão ouro" da *avaliação de impacto*. Com este método, os beneficiários são selecionados aleatoriamente para receber uma intervenção, e cada um tem uma chance igual de receber o programa. Com tamanhos de *amostras* suficientemente grandes o processo de alocação aleatória garante equivalência, tanto em características observadas como não observadas, entre os grupos de tratamento e de controle, resolvendo, assim, qualquer *viés de seleção*.

Amostra aleatória. A melhor maneira de evitar uma *amostra* enviesada ou não representativa é selecionar uma amostra aleatória. Uma amostra aleatória é uma amostra probabilística, em que cada indivíduo da população sendo amostrada tem uma chance (probabilidade) igual de ser selecionado.

Amostra. Em estatística, a amostra é um subconjunto de uma população. Tipicamente, a população é muito grande, tornando um *censo* ou uma enumeração completa de todos os valores na população impraticável ou impossível. Em vez disso, os pesquisadores podem selecionar um subconjunto representativo da população (usando uma base de amostragem) e coletar estatísticas sobre a amostra; estas podem ser usadas para fazer inferências ou extrapolar para a população. Este processo é chamado de *amostragem*.

Amostragem. Processo pelo qual as unidades são retiradas do *quadro de amostragem* construído a partir da *população de interesse* (o universo). Vários procedimentos alternativos de amostragem podem ser utilizados. Métodos de amostragem probabilística são os mais rigorosos porque atribuem uma probabilidade bem definida para cada unidade a ser extraída. A amostragem aleatória, amostragem aleatória estratificada e amostragem por conglomerados são todos métodos de amostragem probabilística.

A amostragem não probabilística (como amostragem intencional ou de conveniência) pode gerar erros de amostragem.

Amostra de conglomerado. Uma *amostra* obtida pela extração de uma *amostra aleatória* de conglomerados (clusters), após a qual ou todas as unidades nos conglomerados selecionados constituem a amostra, ou um número de unidades dentro de cada *conglomerado* selecionado é extraído aleatoriamente. Cada conglomerado tem uma probabilidade bem definida de ser selecionado e cada unidade dentro de um conglomerado selecionado também tem uma probabilidade bem definida de ser extraída.

Amostra estratificada. Obtida dividindo-se a população de interesse (*quadro de amostra*) em grupos (por exemplo, masculino e feminino) e, em seguida, extraindo-se uma *amostra aleatória* de cada grupo. Uma *amostra* estratificada é uma amostra probabilística: cada unidade em cada grupo (ou estrato) tem a mesma probabilidade de ser extraída.

Análise custo-benefício. Cálculos *ex-ante* do total dos custos e benefícios esperados, usados para examinar ou avaliar propostas de projetos. O custo-benefício pode ser calculado *ex-post* em *avaliações de impacto,* se os benefícios puderem ser quantificados em termos monetários e houver informações de custo disponíveis.

Atividade. Ação adotada ou trabalho realizado por meio do qual *insumos*, como recursos financeiros, assistência técnica e outros tipos de recursos são mobilizados para produzir *resultados* específicos.

Atrição. A atrição ocorre quando algumas unidades saem da *amostra* entre uma rodada e outra de coleta de dados, por exemplo, porque os migrantes não são rastreados. A atrição é um caso de *não resposta* da unidade. A atrição pode criar *viés* em avaliações de impacto, se for relacionado à situação de tratamento.

Avaliação de Impacto. Uma *avaliação de impacto* é uma avaliação que tenta estabelecer uma relação causal entre um programa ou intervenção e um conjunto de *resultados*. Uma avaliação de impacto tenta responder à questão de se um programa é responsável por alterações nos resultados de interesse. Contrasta com *avaliação de processo.*

Avaliação de Processo. A *avaliação* de processo é uma avaliação que tenta estabelecer o nível de qualidade ou sucesso dos processos de um programa; por exemplo, a adequação de processos administrativos, a aceitabilidade dos benefícios do programa, a clareza da campanha de informação, a dinâmica interna das organizações implementadoras, seus instrumentos de políticas, seus mecanismos de prestação de serviços, suas práticas de gestão e as ligações entre estes. Contrasta com *avaliação de impacto.*

Avaliação. As avaliações são apreciações objetivas e periódicas de um projeto, programa ou política planejada, em andamento ou concluída. Avaliações são utilizadas para responder a questões específicas, frequentemente relacionadas ao desenho, à implementação e aos resultados.

Cadeia de resultados. A cadeia de resultados define a lógica do programa, que explica como o objetivo de desenvolvimento deve ser alcançado. Ela mostra as vinculações dos *insumos* às atividades, aos *produtos* e aos resultados.

Cálculo de tamanho de amostra Os cálculos de poder estatístico indicam o tamanho da amostra necessário para que uma avaliação detecte um determinado efeito mínimo desejado. Os cálculos de poder estatístico dependem de parâmetros como o poder estatístico (ou a probabilidade de *erro de tipo II*), o *nível de significância*, a variância e a *correlação intraconglomerado* do *resultado* de interesse.

Comparação antes-e-depois. Também conhecida como "comparação *pré-pós*" ou "comparação reflexiva", a comparação antes-e-depois tenta estabelecer o impacto de um programa rastreando mudanças nos *resultados* para os beneficiários do programa ao longo do tempo, através de medições antes e depois do programa, ou política, ser implementada.

Conglomerado (Cluster). Um conglomerado é um grupo de unidades que são semelhantes, de uma maneira ou de outra. Por exemplo, em uma amostra de alunos, crianças que frequentam a mesma escola podem pertencer a um conglomerado porque compartilham as mesmas instalações escolares e professores e vivem no mesmo bairro.

Contrafactual. O contrafactual é uma estimativa do que o *resultado* (*Y*) teria sido para um participante do programa, na ausência do programa (*P*). Por definição, o contrafactual não pode ser observado. Por isso, tem de ser estimado usando *grupos de comparação*.

Correlação de intraconglomerado. Correlação de intraconglomerado é a correlação (ou similaridade) em *resultados* ou características entre as unidades que pertencem ao mesmo conglomerado. Por exemplo, crianças que frequentam a mesma escola normalmente seriam similares ou correlacionadas, em termos da sua área de residência ou nível socioeconômico.

Cumprimento perfeito. No contexto de avaliação de impacto, cumprimento perfeito ocorre quando todos os indivíduos ou unidades para os quais o programa foi oferecido (*grupo de tratamento*) de fato decidem se inscrever no programa, ao passo que todos os indivíduos ou unidades aos quais o programa não foi oferecido (*grupo de comparação*) não participam ou se beneficiam diretamente do programa.

Custo-efetividade. Determinar o custo-efetividade implica comparar intervenções semelhantes com base no custo e na efetividade. Por exemplo, as *avaliações de impacto* de vários programas de educação permitem que os formuladores de políticas tomem decisões mais informadas sobre qual intervenção pode alcançar os objetivos desejados, dado o seu contexto particular e restrições.

Dados da pesquisa. Dados que cobrem uma *amostra* da população de interesse. Contrasta com *dados de censo.*

Dados de censo. Dados que cobrem todas as unidades da população de interesse (*universo*). Contrasta com *dados de pesquisa*.

Diferença-em-diferenças. Também conhecido como "dupla diferença" ou "DD". A diferença-em-diferenças estima o *contrafactual* para a mudança no *resultado* do *grupo de tratamento*, tomando a mudança no resultado do *grupo de comparação*. Este método nos permite levar em conta as diferenças entre os grupos de tratamento e de comparação que permanecem constantes ao longo do tempo. As duas diferenças são, portanto, antes e depois, e entre os grupos de tratamento e de comparação.

Efeito de transbordamento. Também conhecido como contaminação do *grupo de comparação*. O efeito de transbordamento ocorre quando o grupo de comparação é afetado pelo tratamento provido ao *grupo de tratamento*, ainda que o tratamento não seja provido diretamente ao grupo de comparação. Se o efeito de transbordamento no grupo de comparação for negativo (isto é, se são prejudicados pelo programa), então a diferença direta entre os *resultados* no grupo de tratamento e de comparação renderá uma superestimação do impacto do programa. Por outro lado, se o efeito de transbordamento no grupo de comparação for positivo (isto é, eles se beneficiam), então será produzida uma subestimativa do impacto do programa.

Efeito Hawthorne. O "*efeito* Hawthorne" ocorre quando o simples fato de que as unidades estão sendo observadas faz com que se comportem de forma diferente.

Efeito John Henry. O efeito John Henry ocorre quando as unidades de comparação trabalham com mais empenho para compensar não lhes ter sido oferecido o tratamento. Quando se comparam as unidades tratadas a estas unidades de comparação "que se esforçam mais", a estimativa do impacto do programa terá viés; isto é, estimaremos um impacto do programa menor do que o impacto verdadeiro que encontraríamos caso as unidades de comparação não fizessem este esforço adicional.

Efeito. Mudança intencional ou não, devido, direta ou indiretamente, a uma intervenção.

Erro do tipo I. Erro cometido ao rejeitar uma *hipótese nula*, embora a hipótese nula seja válida. No contexto de uma *avaliação de impacto*, o erro do tipo I é cometido quando a *avaliação* conclui que um programa teve um impacto (isto é, a hipótese nula de não impacto é rejeitada) mesmo que, na realidade, o programa não tenha tido impacto (isto é, a hipótese nula é válida). O *nível de significância* determina a probabilidade de um erro do tipo I ser cometido.

Erro do tipo II. Erro cometido não rejeitar uma hipótese nula, embora a hipótese nula seja inválida. No contexto de uma avaliação de impacto, o erro do tipo II é cometido quando a avaliação conclui que um programa não teve impacto (isto é, a hipótese nula de não impacto não é rejeitada), mesmo que o programa tenha tido impacto (isto é, a hipótese nula é inválida). A probabilidade de um erro do tipo II ser cometido é de 1 menos o nível de *poder*.

Estimador Intenção-de-tratar, ou ITT. O *estimador ITT* (do inglês, intention-to-treat) é a diferença direta no *indicador de resultado Y* para o grupo ao qual é oferecido o tratamento e o mesmo indicador para o grupo ao qual não é oferecido o tratamento. Contrasta com o *tratamento-no-tratado*.

Estimador. Em estatística, um estimador é uma estatística (uma função dos dados de *amostra* observáveis) usada para estimar um parâmetro desconhecido da população; uma estimativa é o resultado da aplicação efetiva da função a uma amostra específica de dados.

Grupo de Comparação. Também conhecido como "grupo de controle". Um grupo de comparação válido terá as mesmas características que o grupo de beneficiários do programa (*grupo de tratamento*), exceto que as unidades no grupo de comparação não se beneficiam do programa. Grupos de comparação são usados para estimar o *contrafactual*.

Grupo de Tratamento: Também é conhecido como o grupo tratado ou grupo de intervenção. O grupo de tratamento é o grupo de unidades que se beneficia de uma intervenção, em contraste ao *grupo de comparação,* que não se beneficia.

Hipótese alternativa. Em *avaliação do impacto,* a hipótese alternativa é, geralmente, a hipótese de que a *hipótese nula* seja falsa; em outras palavras, de que a intervenção tem um impacto sobre os *resultados*.

Hipótese nula. Uma *hipótese* nula é a hipótese que poderia ser falseada, com base em dados observados. A hipótese nula, normalmente, propõe uma posição geral ou padrão. Na *avaliação de impacto*, a posição padrão é, em geral, de que não há nenhuma diferença entre os grupos de tratamento e de controle - ou, em outras palavras, de que a intervenção não tem impacto nos *resultados*.

Hipótese. Uma hipótese é uma explicação proposta para um fenômeno observável. Ver também *hipótese nula* e *hipótese alternativa*.

Indicador. Um indicador é uma *variável* que mede um fenômeno de interesse para o avaliador. O fenômeno pode ser um *insumo*, um *produto*, um *resultado*, uma característica ou um atributo.

Insumo. Os recursos financeiros, humanos e materiais utilizados na intervenção de desenvolvimento.

Linha de base. Pré-intervenção, ex-ante. A situação anterior à intervenção, contra a qual pode ser avaliado o progresso ou realizadas comparações. Dados de linha de base são coletados antes que um programa ou política seja implementado, para avaliar a situação "antes".

Método de Regressão Descontínua (RDD). O *Método de Regressão Descontínua* é um método de *avaliação* não experimental. É adequado para programas que utilizam um índice contínuo para classificar os beneficiários potenciais e que têm um limiar ao longo do índice que determina se os potenciais beneficiários receberão o programa ou não. O limite de corte para elegibilidade ao programa oferece um ponto de divisão entre os grupos de *tratamento* e de *comparação*.

Métodos de seleção aleatória. "*Método de seleção aleatória*" é o nome dado ao conjunto de métodos que utilizam a alocação aleatória para identificar o *contrafactual*. Entre eles estão a *alocação aleatória* do tratamento, a *oferta aleatória* do tratamento e a *promoção aleatória*.

Mínimo efeito desejado. A mudança *mínima* em resultados que justificariam o investimento feito em uma intervenção, contando não apenas o custo do programa e os benefícios que este oferece, mas também o custo de oportunidade de não investir fundos em uma intervenção alternativa. O *efeito* mínimo desejado é um insumo para *o cálculo de poder estatístico*, isto é, as *amostras de avaliação* precisam ser grandes o suficiente para detectar pelo menos o efeito mínimo desejado, com *poder estatístico* suficiente.

Monitoramento. O monitoramento é o processo contínuo de coleta e análise de informações para avaliar quão bem determinado projeto, programa ou política, está funcionando. Baseia-se, principalmente, em dados administrativos para rastrear o desempenho (comparando-o com os resultados previstos), fazer comparações entre programas e analisar tendências ao longo do tempo. O monitoramento geralmente rastreia *insumos*, atividades e *produtos*, embora ocasionalmente inclua também *resultados*. O monitoramento é usado para informar a gestão e as decisões diárias.

Não resposta. Dados faltantes ou incompletos de algumas unidades da amostra constituem *não resposta*. A não resposta da unidade surge quando não há informação disponível para algumas unidades da *amostra*, isto é, quando a amostra real é diferente da amostra prevista. *Atrição* é uma forma de não resposta da unidade. Não resposta de item ocorre quando os dados estão incompletos para algumas unidades da amostra em um ponto no tempo. Não resposta pode causar *viés* nos resultados da *avaliação*, se for associada à situação de tratamento.

Nível de significância. O nível de significância é, geralmente, indicado pelo símbolo grego α (alfa). Níveis de significância comuns são de 5% (0,05), 1% (0,01) e 0,1% (0,001). Se um teste de significância der um valor de *p* menor do que o nível α, a *hipótese nula* é rejeitada. Esses resultados são informalmente chamados de "estatisticamente significativos". Quanto mais baixo for o nível de significância, mais forte será a evidência necessária. Escolher o nível de significância é uma tarefa arbitrária mas, para muitas aplicações, o nível de 5 % é escolhido, apenas por ser o convencional.

Oferta Aleatória. Oferta aleatória é um método para identificar o impacto de uma intervenção. Com este método, uma intervenção é oferecida aos beneficiários de forma aleatória e cada um tem uma chance igual de receber o programa. Embora o gestor do programa possa selecionar aleatoriamente as unidades a quem oferecer o tratamento, a partir do universo de unidades elegíveis, o gestor não conseguirá um cumprimento perfeito: ele não pode forçar qualquer unidade a participar ou aceitar o tratamento e não pode recusar-se a deixar que uma unidade participe se a unidade insistir em fazê-lo. No método de oferta aleatória, a oferta aleatória do programa é usada como uma *variável instrumental* para a participação real no programa.

Pareamento. Pareamento é um método de *avaliação* não experimental que utiliza grandes bases de dados e técnicas estatísticas complexas para construir o melhor *grupo de comparação possível* para determinado *grupo de tratamento*.

Pesquisa de seguimento (ou acompanhamento). Também conhecida como pesquisa de "pós-intervenção" ou "ex-post". Uma pesquisa lançada após o início do programa, uma vez que os beneficiários tenham sido beneficiados dele por algum tempo. Uma *avaliação de impacto* pode incluir várias pesquisas de acompanhamento.

Poder estatístico. O poder estatístico é a probabilidade de detectar um impacto se este tiver ocorrido. O poder estatístico de um teste é igual a 1 menos a probabilidade de um erro de tipo II, variando de 0 a 1. Níveis comuns de poder estatístico são 0,8 e 0,9. Níveis altos de poder estatístico são mais conservadores e diminuem a probabilidade de um erro de tipo II. Uma avaliação de impacto tem alto poder estatístico se houver um baixo risco de não detecção dos impactos reais dos programas – ou seja, de ser cometido um erro de tipo II.

População de interesse. O grupo de unidades elegíveis para receber uma intervenção ou tratamento. A população de interesse é, às vezes, chamada de *universo*.

Produto. As entregas, bens de capital e serviços produzidos (fornecidos) diretamente por uma intervenção. Os produtos também podem incluir mudanças resultantes da intervenção e que são relevantes para a obtenção dos *resultados*.

Promoção aleatória. Promoção aleatória é um método semelhante à *oferta aleatória*. No lugar da seleção aleatória das unidades para as quais o tratamento é oferecido, as unidades são selecionadas aleatoriamente para a promoção do tratamento. Deste modo, o programa é deixado aberto para cada unidade.

Quadro de amostragem. A lista mais completa de unidades na *população de interesse* (universo) que pode ser obtida. Diferenças entre o quadro de amostragem e a população de interesse criam um *viés* de cobertura (amostragem). Na presença de *viés* de cobertura, os resultados da *amostra* não terão *validade externa* para toda a população de interesse.

Regressão. Em estatística, a análise de regressão inclui todas as técnicas para modelar e analisar diversas *variáveis,* quando o foco for a relação entre uma variável dependente e uma ou mais variáveis independentes. Na *avaliação de impacto*, a análise de regressão nos ajuda a compreender como o valor típico do *indicador de resultado Y* (variável dependente) se altera quando a alocação ao grupo de tratamento ou ao *grupo de comparação P* (variável independente) variar, enquanto as características dos beneficiários (outras variáveis independentes) são mantidas fixas.

Resultado. Pode ser intermediário ou final. Um resultado é um efeito de interesse que se dá através de uma combinação de fatores de oferta e demanda. Por exemplo, se uma intervenção levar a uma maior oferta de serviços de vacinação, então o número de vacinas aplicadas seria um resultado, uma vez que depende não apenas do fornecimento de vacinas, mas também do comportamento dos beneficiários: eles se apresentam no posto de serviço para ser vacinados? Resultados finais ou de longo prazo são resultados mais distantes. A distância pode ser interpretada em uma dimensão temporal (leva muito tempo para chegar-se ao resultado) ou em uma dimensão causal (muitos nexos causais são necessários para alcançar o resultado).

Tratamento-no-tratado (efeito de). Também conhecido como o *estimador* TOT (do inglês, treatment-on-the-treated). O *efeito* do tratamento no tratado é o impacto do tratamento sobre as unidades que realmente se beneficiaram do tratamento. Contrasta com *intenção-de-tratar*.

Validade externa. Ter validade externa significa que o impacto causal descoberto na *avaliação de impacto* pode ser generalizado para o universo de todas as unidades elegíveis. Para que uma avaliação seja válida externamente, é necessário que a *amostra* de avaliação seja uma amostra representativa do universo de unidades elegíveis.

Validade interna Dizer que uma *avaliação de impacto* tem validade interna significa que ela usa um *grupo de comparação* válido - isto é, um grupo de comparação que é uma estimativa válida do *contrafactual*.

Variável. Na terminologia estatística, uma variável é um símbolo que representa um valor que pode variar.

Variável instrumental. Uma *variável* instrumental é uma variável que ajuda a identificar o impacto causal de um programa quando a participação no programa for parcialmente determinada pelos potenciais beneficiários. Uma variável deve ter duas características para se qualificar como uma boa variável instrumental: (1) deve ser correlacionada com a participação no programa, e (2) não pode ser correlacionada com os *resultados Y* (a não ser através da participação no programa) ou com variáveis não observadas.

Viés de seleção. O viés de seleção ocorre quando as razões pelas quais um indivíduo participa de um programa são correlacionadas com os resultados. Este viés ocorre normalmente quando o *grupo de comparação* é inelegível ou se autoseleciona para fora do tratamento.

Viés. O viés de um *estimador* é a diferença entre a expectativa de um estimador e o verdadeiro valor do parâmetro sendo estimado. Na *avaliação de impacto*, essa é a diferença entre o impacto calculado e o impacto real do programa.

"O objetivo deste livro é fornecer um guia acessível, abrangente e claro para a avaliação de impacto. O material, desde a motivação da avaliação de impacto até as vantagens de diferentes metodologias, cálculos de poder estatístico e de custos, é explicado de forma muito clara e a cobertura é impressionante. Este livro se tornará um guia muito consultado e utilizado e afetará a formulação de políticas para os próximos anos."

Orazio Attanasio, *Professor de Economia, University College London; Diretor, Centro de Avaliação de Políticas de Desenvolvimento, Instituto de Estudos Fiscais, Reino Unido*

"Este é um recurso valioso para quem procura realizar avaliações de impacto no mundo em desenvolvimento, abrangendo os aspectos conceituais e práticos envolvidos, ilustrados com exemplos de prática recente."

Michael Kremer, *Professor de Sociedades em Desenvolvimento, Departamento de Economia, Universidade de Harvard, Estados Unidos*

"Os ingredientes principais para boas avaliações públicas são (a) metodologias apropriadas; (b) a capacidade de resolver problemas práticos, tais como coletar de dados, trabalhar com orçamentos baixos, e escrever relatório final; e (c) governos responsáveis. Este livro não apenas descreve metodologias técnicas sólidas para medir o impacto de programas públicos, mas também fornece vários exemplos e nos leva ao mundo real da implementação de avaliações, desde o convencimento dos formuladores de políticas até a divulgação dos resultados. Se mais profissionais e formuladores de políticas lessem este manual, teríamos melhores políticas e resultados em muitos países. Se os governos melhorarem a prestação de contas, o impacto deste manual será ainda maior."

Gonzalo Hernández Licona, *Secretário Executivo, Conselho Nacional de Avaliação de Políticas de Desenvolvimento Social (CONEVAL), México*

"Eu recomendo este livro como um guia claro e acessível para as questões desafiadoras, práticas e técnicas, enfrentadas na elaboração de avaliações de impacto. Ele se baseia em material que tem sido testado em workshops em todo o mundo e deve se mostrar igualmente útil aos profissionais, formuladores de políticas e avaliadores."

Nick York, *Chefe do Departamento de Avaliação, Departamento para o Desenvolvimento Internacional, Reino Unido*

"O conhecimento é um dos bens mais valiosos para compreender a natureza complexa do processo de desenvolvimento. A avaliação de impacto pode contribuir para preencher a lacuna entre intuição e evidências, para melhor informar a formulação de políticas. Este livro, um dos resultados tangíveis do Fundo Estratégico de Avaliação de Impacto, equipa os profissionais de desenvolvimento humano com ferramentas de ponta para produzir evidências sobre quais políticas funcionam e por quê. Visto que ele aumenta a nossa capacidade de alcançar resultados, esperamos que faça uma grande diferença na prática do desenvolvimento."

Soraya Rodríguez Ramos, *Secretária de Estado para a Cooperação Internacional, Espanha*

ISBN 978-1-4648-0088-7

SKU 210088

GRUPO BANCO MUNDIAL