AULAS 04, 05 E 06
AVALIAÇÃO UTILIZANDO EXPERIMENTOS

Ernesto F. L. Amaral

14, 19 e 21 de março de 2013
Técnicas Avançadas de Avaliação de Políticas Públicas (DCP 098)

Fonte:
Curso "Técnicas Econométricas para Avaliação de Impacto" do "International Policy Centre for Inclusive Growth" (IPC-IG) da "United Nations Development Programme" (UNDP) (http://www.ipc-undp.org/evaluation).

ESTRUTURA DA AULA

– Aleatorização no desenho de pesquisa.

– Análise descritiva das bases de dados.

– Testes de igualdade das médias e das distribuições.

CONTRAFACTUAL

– O problema fundamental da avaliação de impacto é que a unidade de observação (indivíduo, domicílio, município, país) não é observada simultaneamente em dois estados.

– Uma mesma unidade não pode fazer parte do grupo de tratamento e controle ao mesmo tempo (o verdadeiro contrafactual não existe).

– Precisamos então buscar unidades para os dois grupos que sejam o mais semelhantes entre si.

– Porém, as unidades que fazem parte do grupo de tratamento podem ter passado por algum processo seletivo ou auto-seleção.

– Esta seleção diminui a semelhança entre as unidades de tratamento e as unidades de controle.

ALEATORIZAÇÃO NA SOLUÇÃO DE VIÉS DE SELEÇÃO

– O objetivo essencial do trabalho empírico em avaliação de impacto é identificar situações em que se possa assumir que o viés de seleção não existe ou encontrar maneiras de corrigi-lo.

– A aleatorização de indivíduos em grupos de tratamento e controle é um dos casos onde o viés de seleção pode ser inteiramente removido.

– Ou seja, a aleatorização na implementação da política pública (experimento) visa obter grupos de tratamento e de controle que sejam similares em suas características.

GRUPOS DE TRATAMENTO E CONTROLE

– Ao distribuir aleatoriamente as unidades entre os grupos de tratamento (d=1) e controle (d=0), tais unidades diferem apenas quanto a este status, em média.

– Não tivesse o grupo de tratamento sido tratado, ambos os grupos teriam em média o mesmo resultado da variável dependente de interesse.

– Se não houvesse aleatorização, poderia haver viés de seleção das unidades que fariam parte de um grupo de tratamento.

PRESSUPOSTOS NA ALEATORIZAÇÃO

– Não há viés de aleatorização: a aleatorização não afeta o que seria o valor médio da variável de interesse na ausência do tratamento.

– O fato de uma unidade receber o tratamento não afeta o resultado potencial de uma unidade que não recebeu o tratamento (SUTVA).

– Caso o experimento seja implementado de maneira adequada:

  – O efeito médio do tratamento sobre os tratados (ATT), o qual é a diferença entre antes e depois do grupo de tratamento...

  – ... é igual ao efeito médio do tratamento (ATE) na população como um todo (independente de quem foram os grupos de controle e tratamento).

ALEATORIZAÇÃO COMO INSTRUMENTO

– O uso de variável instrumental (IV) tenta resolver o problema de endogeneidade de uma variável explicativa.

– A aleatorização (R) pode ser vista como uma variável instrumental no sentido em que ela determina a participação no programa, mas não é correlacionada com o resultado de interesse.

– Isto é particularmente importante quando os indivíduos decidem participar ou não em um programa, baseados em suas expectativas de ganho que podem ser uma função tanto de variáveis observáveis (X) como de variáveis não observáveis (U).

MAIS PRESSUPOSTOS

– Nestas circunstâncias, o efeito médio do tratamento sobre os tratados (ATT) será igual ao efeito médio do tratamentos (ATE), sob algumas condições.

– Não há heterogeneidade do impacto:

  – Impacto é o mesmo para pessoas com as mesmas características X. O componente não observável não afeta o ganho.

– Há controle dos componentes não observáveis:

  – Há diferença entre os componentes não observáveis do tratamento e controle ($U_1 \neq U_0$), mas a aleatorização faz com que tais características sejam semelhantes entre os grupos de tratamento e controle.

DESENHANDO UM PLANO AMOSTRAL

– Ao selecionar as unidades dos grupos de tratamento e controle que fornecerão informações para a avaliação da política pública, é preciso elaborar um plano amostral:

  – Tamanho da amostra.
  – Erros agrupados (*cluster*).
  – Alocação imperfeita (*imperfect compliance*).
  – Variáveis de controle.
  – Estratificação.
  – Poder de teste.

– Quanto mais complexo for o plano amostral, maior a necessidade de ter o apoio de um estatístico em sua formulação.

TAMANHO DA AMOSTRA

– A amostra precisa ter um tamanho suficiente que permita identificar um impacto de tamanho x.

– A proporção ótima é ter um grupo de comparação maior, já que as unidades neste grupo são mais heterogêneas.

MARGEM DE ERRO

– Quando coletamos um conjunto de dados amostrais, podemos calcular a proporção amostral, a qual é tipicamente diferente da proporção populacional.

– A **margem de erro** (***E***) é a diferença máxima provável entre a proporção amostral observada e o verdadeiro valor da proporção populacional:

  – Isso ocorre quando dados de amostra aleatória simples são usados para estimar uma proporção populacional.
  – É também chamada de erro máximo da estimativa.
  – É encontrada pela multiplicação do valor crítico pelo desvio padrão das proporções amostrais.

MARGEM DE ERRO E INTERVALO DE CONFIANÇA

– Margem de erro para proporções é calculada por:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

– Há uma probabilidade *α* de que a proporção amostral tenha erro maior do que *E*.

– Ou seja, $\hat{p}$ terá probabilidade de 1 – *α* de estar a: $z_{\alpha/2}\sqrt{pq/n}$ de *p*.

– Intervalo de confiança para proporção populacional é representado por: $\hat{p} - E < p < \hat{p} + E$

$$\hat{p} \pm E$$

$$(\hat{p} - E; \hat{p} + E)$$

CONSTRUÇÃO DE INTERVALO DE CONFIANÇA

– Verifique se requisitos são satisfeitos: (1) amostra aleatória simples; (2) condições para distribuição binomial (tentativas fixas, independentes, duas categorias, probabilidade constante); e (3) há pelo menos 5 sucessos e 5 fracassos.

– Ache o valor crítico que corresponde ao nível de confiança desejado. Se nível de confiança é 95%, $z_{\alpha/2} = 1{,}96$.

– Calcule a margem de erro:

$$E = z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$$

– Use o valor da margem de erro e o valor da proporção amostral para encontrar o intervalo de confiança:

$$\hat{p} - E < p < \hat{p} + E$$

– Arredonde os limites do intervalo de confiança.

COMO DEFINIR O TAMANHO AMOSTRAL?

– Utilizando a fórmula da margem de erro, chegamos a:

$$n = \frac{\left[z_{\alpha/2}\right]^2 \hat{p}\hat{q}}{E^2}$$

– Se não conhecemos qualquer estimativa $\hat{p}$:

$$n = \frac{\left[z_{\alpha/2}\right]^2 0{,}5 * 0{,}5}{E^2} = \frac{\left[z_{\alpha/2}\right]^2 0{,}25}{E^2}$$

– Se o tamanho amostral calculado não for um número inteiro, arredonde-o para o inteiro maior mais próximo.

– Quando a amostragem é sem reposição, a partir de uma população finita relativamente pequena, utilize:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\sqrt{\frac{N-n}{N-1}} \qquad n = \frac{\left[z_{\alpha/2}\right]^2 N\hat{p}\hat{q}}{\hat{p}\hat{q}\left[z_{\alpha/2}\right]^2 + (N-1)E^2}$$

TAMANHO DA POPULAÇÃO

– Para o cálculo do tamanho da amostra, o tamanho da população é usado somente em casos em que fazemos amostragem sem reposição a partir de uma população relativamente pequena.

– **Outras observações:**

 – Se margem de erro desejada igual a 5%, E=0,05.

 – Se nível de confiança desejada é de 95%, $z_{\alpha/2}$=1,96.

 – Assim:

$$n = \frac{\left[z_{\alpha/2}\right]^2 0{,}25}{E^2} = \frac{(1{,}96)^2 * 0{,}25}{(0{,}05)^2} = \frac{0{,}9604}{0{,}025} = 384{,}16 \approx 385$$

MARGEM DE ERRO E INTERVALO DE CONFIANÇA

– Para calcular margem de erro *E* para estimativa de *μ* com *σ* desconhecido, onde $t_{\alpha/2}$ tem *n*–1 graus de liberdade:

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

– Intervalo de confiança para estimativa de *μ* com *σ* desconhecido:

$$\bar{x} - E < \mu < \bar{x} + E$$

CONSTRUÇÃO DE INTERVALO DE CONFIANÇA

– Verifique se os requisitos são satisfeitos: (1) amostra aleatória simples; e (2) população próxima de distribuição normal ou n>30.

– Usando n–1 graus de liberdade, ache valor crítico $t_{\alpha/2}$, correspondente ao nível de confiança.

– Calcule margem de erro: $E = t_{\alpha/2}\, s/\sqrt{n}$

– Use valor da margem de erro e valor da média amostral e ache os valores dos limites do intervalo de confiança:

$$\bar{x} - E < \mu < \bar{x} + E$$

– Arredonde os limites do intervalo de confiança resultante.

ERROS AGRUPADOS (*CLUSTER*)

– Quando se aleatoriza grupos ao invés de indivíduos, é importante observar que o erro pode não ser independente entre indivíduos dentro do mesmo grupo.

– Neste caso, a precisão das estimativas aumenta mais com um aumento do número de grupos (*clusters*) do que com um aumento nas observações (n) dos grupos existentes.

ALOCAÇÃO IMPERFEITA

– Avaliar se houve alocação imperfeita (*imperfect compliance*).

– Pode haver unidades de observação que não receberam política dentro do grupo de tratamento.

– Pode haver unidades que receberam política no grupo de controle.

– Isso exige um tamanho da amostra maior, já que a exclusão das unidades mal identificadas pode diminuir a significância estatística dos resultados obtidos.

– Se não for possível aumentar amostra, pode-se estimar a estimativa da intenção do tratamento (ITT): diferença entre grupo de tratamento e controle, segundo informações do questionário e não da moldura da amostragem (lista de unidades da qual a amostra é selecionada).

VARIÁVEIS DE CONTROLE

– Controlar por variáveis que influenciam a variável de interesse não afeta o valor esperado do estimador (β), mas pode reduzir sua variância e aumentar sua significância estatística, o que diminui o tamanho necessário da amostra.

– As variáveis de controle devem ser capturadas no momento anterior à implementação da política (*baseline*), pois se elas forem influenciadas pelo tratamento, elas podem capturar parte do efeito deste tratamento na variável de interesse.

– Controlando por variáveis que afetam pouco a variável de interesse pode aumentar a variância do estimador (β).

– Se há problemas na aleatorização, há ameaça à validade interna da avaliação, o que pode ser equacionado com variáveis de controle, incluindo interações.

ESTRATIFICAÇÃO

– A estratificação é usada para garantir que para certas dimensões observáveis (X), os grupos de controle e tratamento tenham médias (esperanças) semelhantes.

– Estratificação é utilizada para garantir que isso se verifique na prática.

– A precisão será melhorada na medida em que os blocos forem formados por variáveis que afetam a variável de interesse.

– Além de reduzir a variância do estimador (β), este procedimento é útil para analisar a heterogeneidade do impacto para diferentes grupos.

PODER DE TESTE

– Usamos β para designar a probabilidade de deixar de rejeitar uma hipótese nula falsa (**erro tipo II**).

– **Poder de um teste** de hipótese é a probabilidade (κ=1–β) de se rejeitar uma hipótese nula falsa.

  – Essa probabilidade é calculada usando um **nível de significância** específico (α) e um valor particular do parâmetro populacional que seja uma alternativa (H_1) ao valor assumido na hipótese nula (H_0).

– O **poder de um teste** de hipótese é a probabilidade de se apoiar uma hipótese alternativa (H_1) verdadeira.

– **Dependendo dos valores** particulares escolhidos como alternativos à hipótese nula, poder do teste será diferente.

– Geralmente é exigido poder de teste entre 0,8 e 0,9.

CÁLCULO DO PODER DE TESTE NA PRÁTICA

– É necessário ter uma idéia da média e da variância da variável de interesse na ausência do experimento, depois de controlar por possíveis covariáveis e/ou estratificação.

– No caso de desenho amostral agrupado (*cluster*) é preciso ter uma idéia da correlação da variável de interesse para membros do mesmo grupo.

– O STATA pode auxiliar no cálculo do tamanho da amostra, ao estabelecer um poder de teste específico.

TAMANHO DA AMOSTRA E PODER DE TESTE NO STATA

– Utilize o comando:

sampsi #1 #2, sd(#) alpha(#) power(#) n(#) onesamp

– *#1*: média na população (hipótese nula).

– *#2*: média alternativa (hipótese alternativa).

– *sd*: desvio padrão da população.

– *alpha*: nível de significância adotado.

– *power*: poder de teste.

– *n*: tamanho da amostra.

– *onesamp*: teste de uma amostra.

DEFININDO TAMANHO DA AMOSTRA (*n*)

– Uma pesquisa verificou 40% de intenção de voto no candidato A, com desvio padrão de 10%. Hipótese alternativa é que a pesquisa subestimou intenção de voto em 5%. Qual o tamanho da amostra a ser coletada para que H_1 seja provada com margem confiável?

– H_0: V_A=40%

– H_1: V_A=45%

– Desvio padrão=10%

– Utilizamos: α=0,05 (prob. rejeitar H_0 quando é verdadeira)

– Utilizamos: (1–β)=0,90 (prob. rejeitar uma H_0 falsa).

sampsi #1 #2, sd(#) alpha(#) power(#) n(#) onesamp

sampsi 40 45, sd(10) alpha(.05) power(.9) onesamp

RESULTADO DO TAMANHO DA AMOSTRA (*n*)

```
. sampsi 40 45, sd(10) alpha(0.05) power(0.9) onesamp

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =     40, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.9000
 alternative m =       45
            sd =       10

Estimated required sample size:

             n =       43
```

– Quanto maior desvio padrão, maior *n*.

– Quanto maior *α*, menor nível de confiança (1–α), menor *n*.

– Quanto maior poder de teste (1–*β*), maior *n*.

– Quanto maior diferença entre H_0 e H_1, menor *n*.

INTERPRETAÇÃO DO TAMANHO DA AMOSTRA (*n*)

– O resultado indica que, com nível de significância de 0,05 e poder de teste de 90%, seriam necessárias 43 entrevistas selecionadas aleatoriamente para detectar um aumento da intenção de voto no candidato A de 40% para 45%.

DEFININDO PODER DE TESTE (1–β)

- Uma pesquisa verificou 40% de intenção de voto no candidato A, com desvio padrão de 10%. Hipótese alternativa é que a pesquisa subestimou intenção de voto em 5%. Se testamos essa pesquisa com uma amostra de tamanho 20, qual o poder de teste neste caso?
- H_0: V_A=40%
- H_1: V_A=45%
- Desvio padrão=10%
- Tamanho da amostra (*n*)=20
- Utilizamos: α=0,05 (prob. rejeitar H_0 quando é verdadeira)

sampsi #1 #2, sd(#) alpha(#) power(#) n(#) onesamp

sampsi 40 45, sd(10) alpha(.05) n(20) onesamp

RESULTADO DO PODER DE TESTE (1–β)

```
. sampsi 40 45, sd(10) alpha(.05) n(20) onesamp

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =     40, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (two-sided)
 alternative m =       45
            sd =       10
 sample size n =       20

Estimated power:

         power =   0.6088
```

– Quanto maior desvio padrão, menor poder de teste (1–β).

– Quanto maior α, menor β, maior poder de teste (1–β).

– Quanto maior n, maior poder de teste.

– Quanto maior diferença entre H_0 e H_1, maior poder de teste.

INTERPRETAÇÃO DO PODER DE TESTE (1–β)

– O resultado indica que, com nível de significância de 0,05 e 20 entrevistas selecionadas aleatoriamente, a pesquisa teria um poder de teste de 61% para detectar um aumento da intenção de voto no candidato A de 40% para 45%.

IMPLEMENTANDO A ALEATORIZAÇÃO

– Ao implementar a aleatorização na implementação de uma política pública, é preciso pensar em:

  – Nível da aleatorização.

  – Desenhos fatoriais (*cross-cutting design*).

  – Levantamento de dados.

NÍVEL DA ALEATORIZAÇÃO

– Em alguns casos é obvio se o nível de aleatorização deve ser o indivíduo ou algum grupo (escola, cidade, indústrias), mas em outros casos não é tão óbvio assim.

– Critérios para decisão:
  - – Quanto maior o número de grupos, maior o tamanho da amostra para o tamanho mínimo do efeito detectável (MDE).
  - – Efeito de transbordamento (ou externalidade) da política pode enviesar a estimação dos efeitos do tratamento.
  - – Impactos nos indivíduos não selecionados e custos fixos.

DESENHOS FATORIAIS (*CROSS-CUTTING DESIGN*)

– Usados para testar várias intervenções e combinações destas em relação a um grupo de comparação e entre grupos de tratamento.

– Usados para testar interações de diferentes componentes de um programa.

– Pode ser usado para testar duas hipóteses ao invés de uma, sem um grande aumento de custos.

LEVANTAMENTO DE DADOS

– Linha de base (marco zero, *baseline*):
  – Reduz os requisitos do tamanho da amostra, por gerar variáveis de controle correlacionadas com a variável de interesse.
  – Interação entre valores iniciais e o impacto do programa.
  – Checar se a aleatorização foi bem feita.
  – Testar e refinar os procedimentos de coleta de dados.

– Uso de dados administrativos:
  – Metodologia de coleta quando se combina pesquisa de campo e dados administrativos.