

ALEATORIZAÇÃO



Objetivo

- Discutir o conceito, vantagens e limitações do uso da aleatorização como ferramenta de estabelecimento de relação de causalidade em economia



Bibliografia

- Banerjee, A.; Duflo, E. “The experimental approach do development economics”, NBER Working Paper, 14467, 2008
- Duflo, E.; Glennerster, R.; Kremer, M. “Using Randomization in Development Economics Research: A Toolkit”, Handbook of Development Economics, 2007
- Imbens e Wooldridge, JEL, 2009



Bibliografia

- Levitt, S.; List, J. “Field experiments in economics: The past, the present, and the future”, European Economic Review, 2009
- List, Sadoff, Wagner, “So you want to run an experiment, now what? Some Simple Rules of Thumb for Optimal Experimental Design”, NBER, 2010



INTRODUÇÃO

- Na bioestatística, análise experimental é vista como único meio crível de estabelecimento de causalidade (Imbens e Wooldridge, 2009).
 - Ex: the US Food and Drug Administration requires evidence from randomized experiments in order to approve new drugs and medical procedures.
- Importante nessas análises: garantir que não haja interações entre participantes e não participantes para que se possa estabelecer relação de causalidade, o que é mais fácil em medicina do que em economia.



INTRODUÇÃO

- Causalidade é estabelecida pois qualquer diferença sistemática entre grupos de tratamento e controle passa a ser resultado direto da intervenção.
- Assim, os possíveis vieses do estimador “naive” são eliminados.
- Apesar de simplificar a avaliação do efeito do programa, a implantação da aleatorização pode sofrer diversas dificuldades e há diversas escolhas a serem feitas.



ALEATORIZAÇÃO - QUESTÕES

- Quem deve fazer a aleatorização (governo, ONG)?
- Como aleatorizar sem criar problemas éticos e políticos?
- Como determinar o tamanho amostral mínimo para que se possa rejeitar a hipótese nula que não há efeito?
- Em que nível aleatorizar (cidades, bairros, indivíduos)?
- Deve-se estratificar? Quando coletar dados?



PROGRESA (MÉXICO)

- É provavelmente o mais conhecido exemplo de uma avaliação randomizada conduzida pelo governo em países em desenvolvimento.
- Programa foi lançado em 1998. Como restrição orçamentária impossibilitava alcançar 50 mil potenciais comunidades de uma vez, a opção foi começar com programa piloto randomizado em 506 comunidades.



PROGRESA (MÉXICO)

- Metade das comunidades foi aleatoriamente selecionada para receber o programa;
- Dados de *baseline* e dados de *follow-up* foram coletados para as comunidades dos dois grupos.
- Diferentes pesquisadores tiveram acesso aos dados e vários papers foram escritos sobre o impacto do programa (p. ex. Behrman e Hoddinott (2005), impacto em nutrição infantil.)
- Programa foi efetivo em melhorar saúde e educação.



OPORTUNIDADES PARA ALEATORIZAÇÃO

1. Uma janela natural para introduzir randomização é no momento anterior ao programa ser expandido, durante a fase piloto, como no Progresas.
- Nesse aspecto tem crescido estudos onde as agências de implementação dos programas (não necessariamente governo) e os pesquisadores transformam a avaliação em um “experimento de campo”, no sentido que ambos querem encontrar a melhor solução para um problema.



STEPS (2016)

- Intervenção era para ser aleatorizada, mas acabou não sendo...
- Dois tratamentos:
 - Intervenção semanal com equipe STEPS e auxílio das professoras
 - Intervenção semanal com as professoras e supervisão mensal da equipe STEPS



OPORTUNIDADES PARA ALEATORIZAÇÃO

2. Situações onde há restrições de orçamento ou de capacidade de implementação e a demanda pelo programa excede a oferta.
- Um caminho natural para ‘racionar’ os recursos é selecionar por sorteio, entre os candidatos elegíveis, aqueles que receberão o programa.



EXEMPLO: LIGA SOLIDÁRIA (2015)

- Programa de qualificação profissional de jovens em SP.
- Mais inscritos do que vagas em alguns dos cursos
- Após realização das inscrições e provas de conhecimentos básicos de português e matemática, além de exclusão dos que não pertenciam ao público-alvo → sorteio



OPORTUNIDADES PARA ALEATORIZAÇÃO

3. Restrições financeiras e administrativas levam as ONG's a introduzir o programa aos poucos (como no exemplo com o Progres). Randomização é a forma mais justa de decidir quem será contemplado primeiro.
- Aleatorizar a ordem do tratamento permite a avaliação do programa em contextos onde não seria aceitável para alguns grupos ou indivíduos ficarem sem suporte.



PROBLEMAS

- Dificuldade para medir efeitos de longo prazo;
- Se a aleatorização *phase-in* for rápida demais relativamente ao tempo que o programa demora a surtir efeito, não é possível detectar o efeito do tratamento.
- Grupo de comparação é afetado pela expectativa futura de tratamento.



EXEMPLO

- Programa de Microcrédito: indivíduos no grupo de controle podem atrasar seu investimento antecipando que terão crédito mais barato uma vez que tiverem acesso ao programa de microcrédito no futuro.
- Deixa de ser um bom contra-factual, pois também é afetado pelo tratamento.



PROBLEMA DO PHASE-IN

- Em alguns casos, as pessoas do grupo de controle podem não topar participar da coleta de dados enquanto não estiverem no programa.
- Possível saída: within-group randomization



PROBLEMA DO PHASE-IN

- Banerjee et al. 2007 avaliam efeito de reforço escolar. Ao invés de dar o tratamento no 3^o e no 4^o ano em uma mesma escola, dá o tratamento no 3^o ano de uma escola e o de 4^o ano na outra escola.
- Problema: possível contaminação



ALEATORIZAÇÃO DO ENCORAJAMENTO

- Em situações em que não é possível ou ético realizar a aleatorização, uma saída é aleatorizar o encorajamento para o recebimento do tratamento.
- Ex: envio aleatoriamente material publicitário falando das vantagens de se fazer um curso de qualificação profissional.
- Isso deve aumentar a chance de cursar e serve como instrumento para o efeito de ter feito o curso sobre o resultado no mercado de trabalho.



TAMANHO DA AMOSTRA NA ALEATORIZAÇÃO

- Uma das primeiras perguntas em uma aleatorização é: Qual o tamanho de amostra que vamos precisar?
- Temos que pensar no poder do teste: qual a probabilidade de que para um dado tamanho de efeito e nível de significância, será possível rejeitar a hipótese nula de que não há efeito do tratamento?



Erros tipo I e tipo II

		Intervenção tem efeito positivo (desconhecido)	
		Não (H_0 verdadeira)	Sim (H_0 falsa)
Ação	Rejeitamos H_0	ERRO TIPO I = α	(OK) $1-\beta$
	Não rejeitamos H_0	(OK) $1-\alpha$	ERRO TIPO II = β



Erro tipo I e Erro tipo II

- Erro tipo I = α
- Probabilidade de rejeitar H_0 quando esta é verdadeira
- **Probabilidade de concluir que a intervenção tem efeito quando o efeito é nulo**
- Valores típicos: 0.01, 0.05, 0.1



Erro tipo I e Erro tipo II

- Erro tipo II = β
 - Prob. de não rejeitar H_0 quando ela é falsa
- Poder Estatístico = $1 - \beta$
 - Probabilidade de rejeitar H_0 quando H_0 é falsa
- **Probabilidade de concluir que a intervenção tem efeito toda vez que ela realmente tiver efeito. Valor típico: 0.80**



PODER

- Em um estudo comparando dois grupos, o poder é a probabilidade de rejeitar a hipótese nula de que os dois grupos tenham a mesma média populacional e, portanto, concluir que existe diferença de uma dada magnitude.
- É a probabilidade de decidir corretamente de que os dois grupos são diferentes.



TESTE DE HIPÓTESES

- Seja $Y_i = \alpha + \gamma T_i + u_i$
- Onde T_i é uma variável binária de tratamento e Y_i é a variável de resultado de interesse. Seja P a proporção de pessoas que recebem o tratamento.
- Em MQO, $var(\hat{\gamma}) = \frac{\sigma^2}{\sum_{i=1}^n (T_i - \bar{T})^2} =$



TESTE DE HIPÓTESES

$$\begin{aligned}
 \bullet \text{ var}(\hat{\gamma}) &= \frac{\sigma^2}{\sum_{i=1}^n (T_i - \bar{T})^2} = \frac{\sigma^2}{\sum_{i=1}^n (T_i^2 - 2T_i\bar{T} + \bar{T}^2)} = \\
 &= \frac{\sigma^2}{\sum_{i=1}^n (T_i^2 - 2T_i\bar{T} + \bar{T}^2)} = \\
 &= \frac{\sum_{i=1}^n T_i^2 - 2\bar{T} \sum_{i=1}^n T_i + n\bar{T}^2}{\sigma^2} = \\
 &= \frac{\sum_{i=1}^n T_i^2 - 2\bar{T} \sum_{i=1}^n T_i + n\bar{T}^2}{\sigma^2} = \frac{\sum_{i=1}^n T_i^2 - 2n\bar{T} \sum_{i=1}^n T_i + n\bar{T}^2}{\sigma^2}
 \end{aligned}$$



TESTE DE HIPÓTESES

$$\sigma^2 = \frac{\sigma^2}{\sum_{I=1}^n T_i^2 - n\bar{T}^2} =$$

Note: $\sum_{I=1}^n T_i^2 = \sum_{I=1}^n T_i$, pois T_i é binário (0/1)

$$\frac{\sigma^2}{\sum_{I=1}^n T_i - n\bar{T}^2} = \frac{\sigma^2}{n\bar{T} - n\bar{T}^2}. \text{ Portanto:}$$

$$\text{var}(\hat{\gamma}) = \frac{\sigma^2}{n\bar{T}(1-\bar{T})} = \frac{\sigma^2}{nP(1-P)}$$

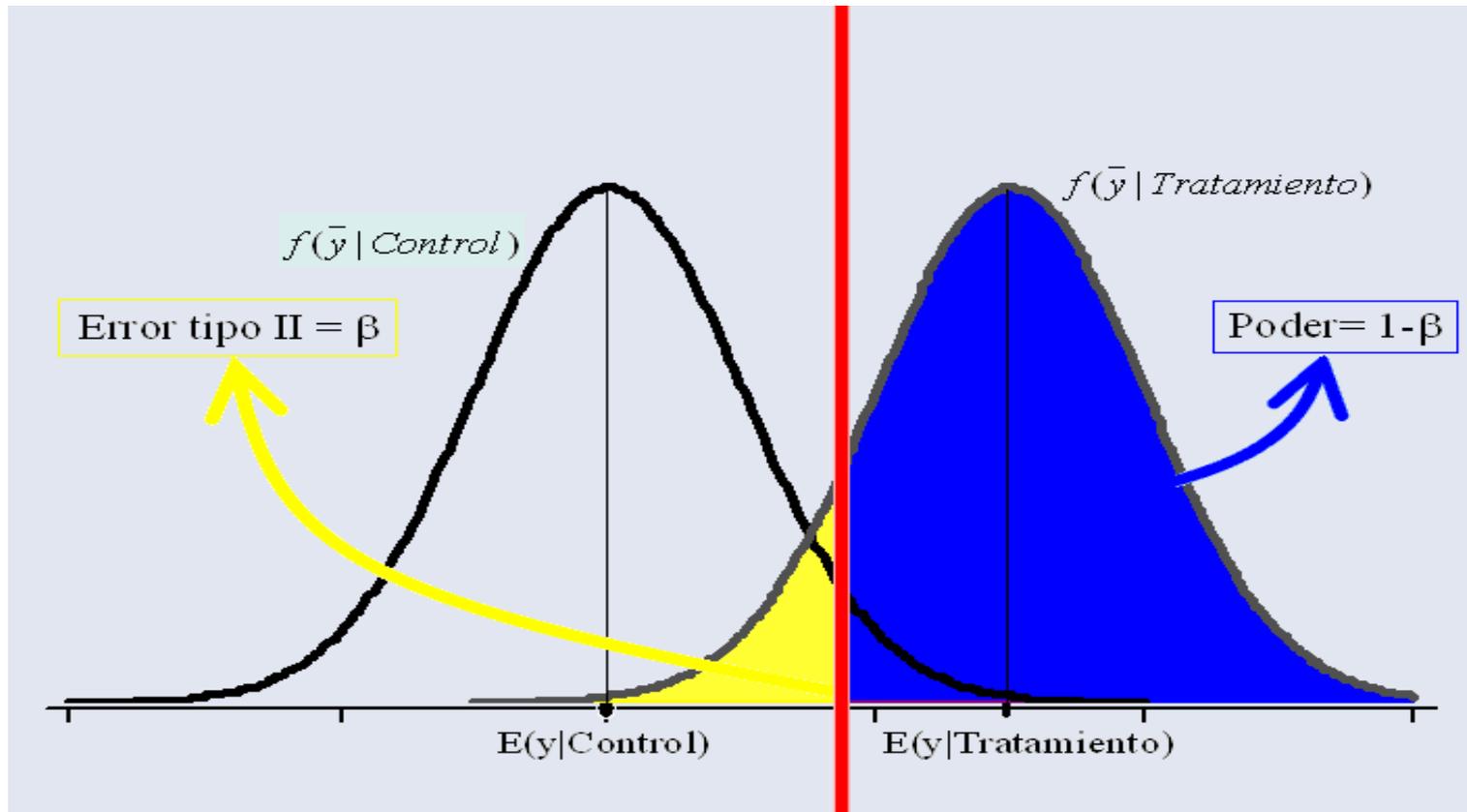


Poder do Teste

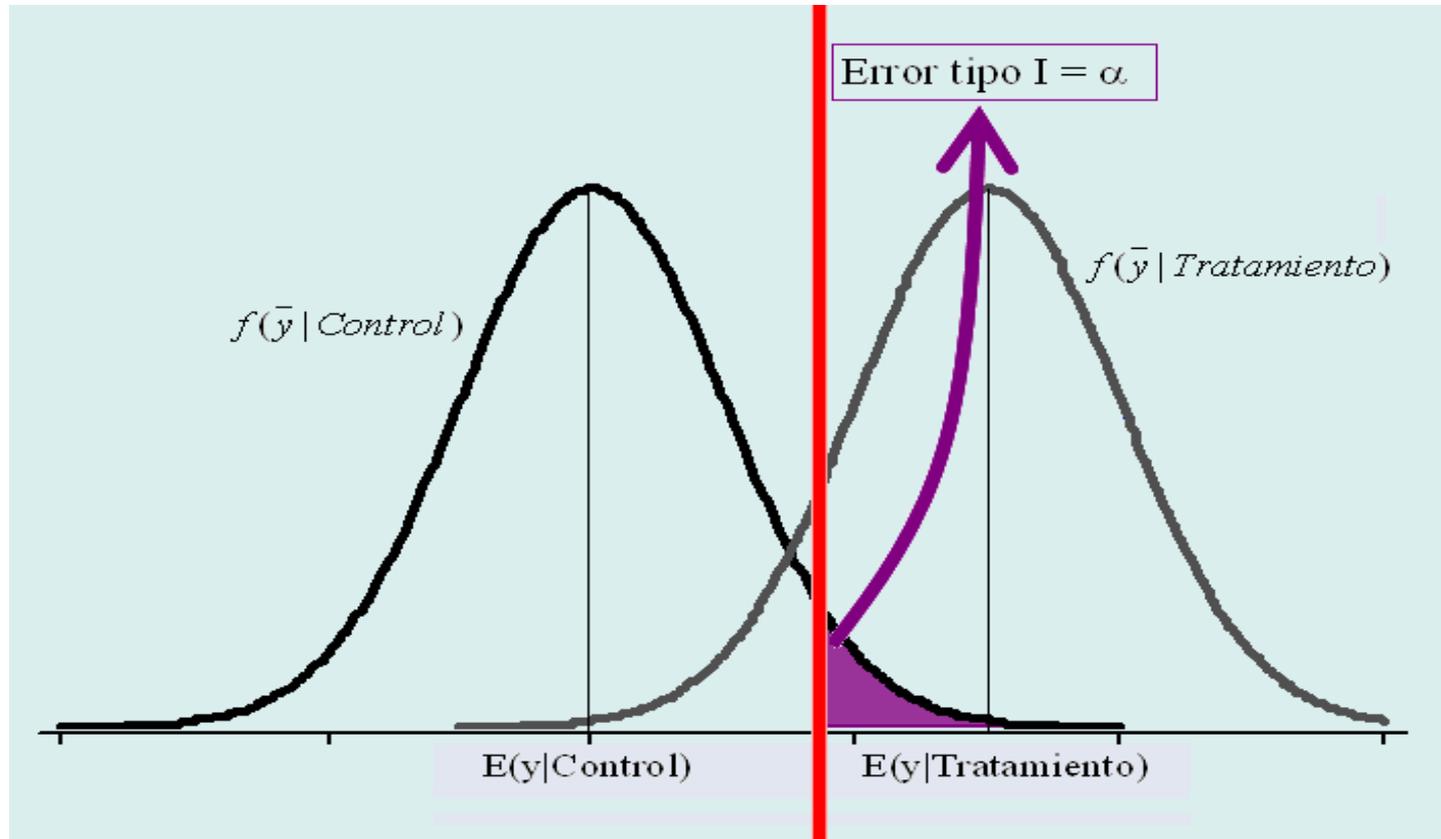
- Qual a Probabilidade de rejeitar $H_0: \gamma = 0$ quando na verdade $\gamma \neq 0$?
- R: Poder do teste



Erro de tipo II e poder do teste



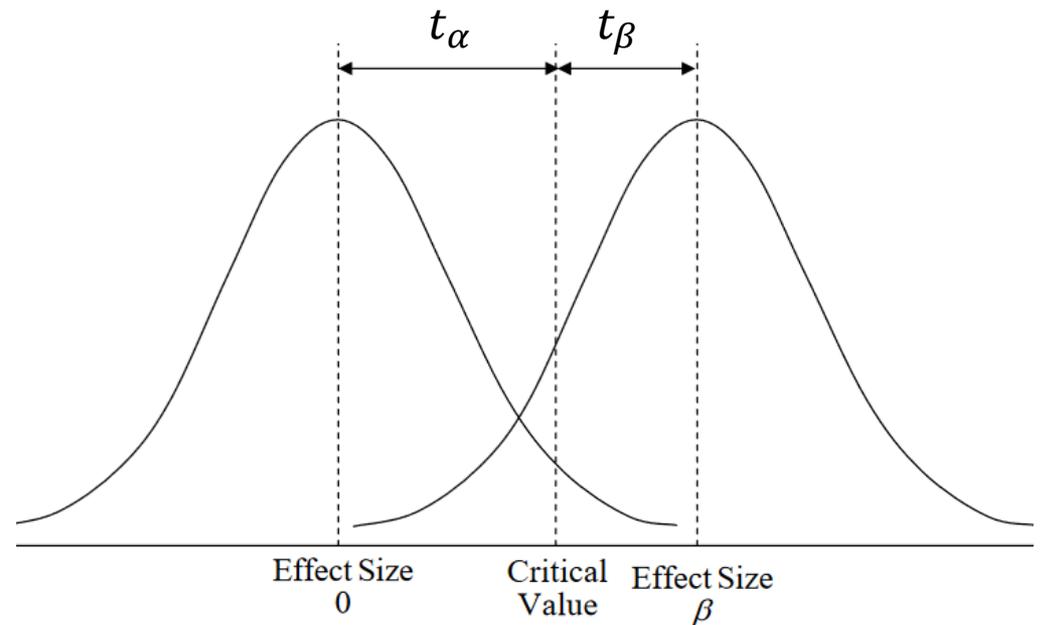
Erro de tipo I



Poder do Teste

- Para calcular o poder do teste, deve valer:

- $\gamma > (t_\beta + t_\alpha)ep(\hat{\gamma})$



Efeito Mínimo Detectável (EMD)

- EMD para um poder β , nível de significância α , tamanho de amostra n , e fração de tratados P é dado por:

$$EMD = (t_\beta + t_c)ep(\hat{\gamma}) =$$

$$= (t_\beta + t_c) \sqrt{\frac{\sigma^2}{nP(1-P)}}$$



Efeito Mínimo Detectável (EMD)

- EMD para um dado poder $(1 - \beta)$, nível de significância α e fração de tratados P é:

- $$EMD = (t_\beta + t_\alpha) \sqrt{\frac{\sigma^2}{nP(1-P)}}$$

- EMD diminui com redução do poder ($\downarrow t_\beta$),
- Aumento de α ($\downarrow t_\alpha$); Redução de σ^2
- Aumento do tamanho da amostra



Efeito Mínimo Detectável (EMD)

- Como a distribuição de t sob alternativa não é centrada em zero, essa conta não é correta para valores pequenos de amostra.
- Assim, dever ser usada apenas quando a t -student converge para a $N(0,1)$.
- Para n um pouco menores, dever ser usada a distribuição correta da T não centrada.



T não centrada

- Se Z é $N(0,1)$ e V é qui-quadrada com v graus de liberdade, independente de Z , define-se a distribuição t não centrada como:

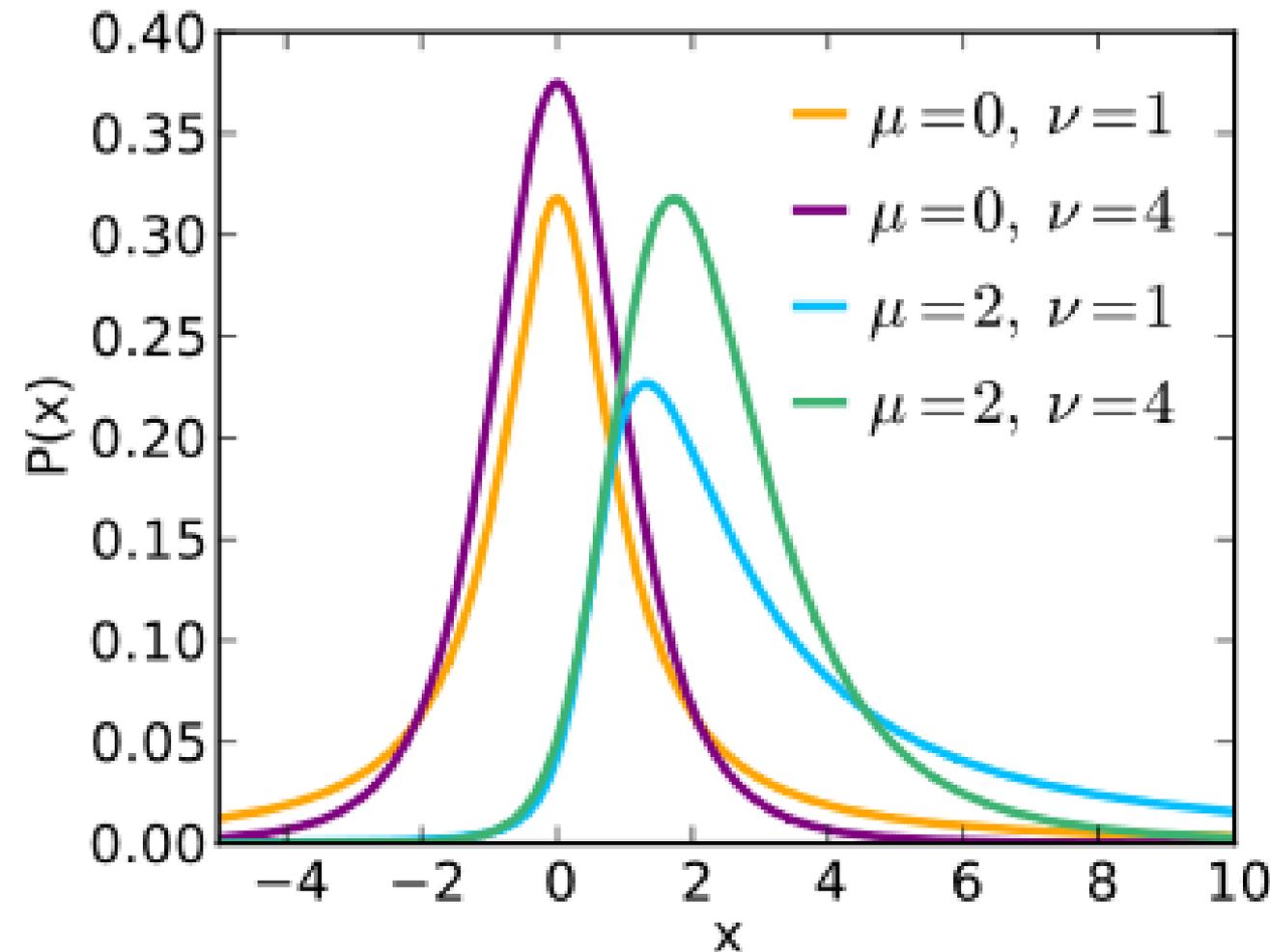
- $$t(v, \mu) = \frac{Z + \mu}{\sqrt{\frac{V}{v}}}$$

- Onde v são os graus de liberdade e μ é o parâmetro de centralidade



T não centrada

- Com 1 g.l., a t não centrada em 0 é bem assimétrica.
- Com 4 g.l., ela fica mais parecida com a centrada



T não centrada

2.1 One-sample test of mean

Suppose that we have a single sample, x_i , $i = 1, \dots, n$, which we assume comes from a normal distribution with mean μ and standard deviation σ . We wish to test the hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_a : \mu \neq \mu_0$$

for some hypothesized value μ_0 .

The standard parametric test for this situation is the one-sample t test. This test is based on the test statistic

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where \bar{x} represents the sample mean and s the sample standard deviation.

Fonte: Harrison e Brady (2004)



T não centrada

Under the null hypothesis, T has a Student's t -distribution on $n - 1$ degrees of freedom. Under the alternative hypothesis, T has a noncentral t -distribution on $n - 1$ degrees of freedom with noncentrality parameter

$$\theta = \frac{\mu - \mu_0}{\sigma / \sqrt{n}}$$

The power to detect a difference of $\delta = \mu - \mu_0$ with two-sided significance level α is given by

$$1 - \beta = T_{n-1} \left(t_{\alpha/2, n-1} \left| \frac{\delta \sqrt{n}}{\sigma} \right. \right) - T_{n-1} \left(-t_{\alpha/2, n-1} \left| \frac{\delta \sqrt{n}}{\sigma} \right. \right)$$

where $T_{df}(\cdot | \theta)$ is the cumulative distribution function of the noncentral t -distribution with df degrees of freedom and noncentrality parameter θ and $t_{p, df}$ is the point of the central t -distribution with df degrees of freedom corresponding to an upper-tail probability of p .



DUAS ABORDAGENS:

- **Abordagem da determinação do poder:**
- Admite-se um certo efeito mínimo detectável e um nível de poder com o objetivo de descobrir qual tamanho de amostra teremos.



ABORDAGEM DA DETERMINAÇÃO DO PODER

- Ex: Pesquisador quer avaliar efeito de intervenção em nível de escola para melhorar a nota. Aleatorizam as escolas e teoria sugere que, na prática, um efeito significativo giraria em torno de 0,20 desvios padrão.
- Nesse caso, se atribuo poder de 80%, vou achar o tamanho de amostra necessário para tanto.



ABORDAGEM DO TAMANHO DO EFEITO

- Definido um nível de poder e um tamanho de amostra , achar qual é o EMD.
- Ex: Por restrições financeiras, só pode aplicar em 50 escolas e 100 alunos em cada. Assim, dado o tamanho da amostra, vou procurar qual o menor efeito detectável com essa amostra.



PASSOS DE AVALIAÇÃO ALEATORIZADA

1. Selecionar aleatoriamente um grupo de indivíduos de uma população bem definida
2. Esses indivíduos são alocados aleatoriamente ao grupo de tratamento ou de controle
3. Constroi-se uma linha de base e checa observáveis, se aleatorização foi bem feita
4. Uma intervenção pré-definida é aplicada apenas ao grupo de tratamento



PASSOS DE AVALIAÇÃO ALEATORIZADA

5. Resultados são medidos para cada participante
 6. A diferença das médias dos dois grupos, se for estatisticamente significativa, indica que houve efeito da política e corresponde ao ATE, que nesse caso será igual ao ATT e ATNT
- A razão pela qual aleatorização garante estimativa não viesada é porque todas as características, mesmo as não observáveis, serão distribuídas igualmente nos grupos.



PASSOS DE AVALIAÇÃO ALEATORIZADA

- Ou seja, o contra-factual, nesse caso, é representado pelos indivíduos que foram sorteados para ficar no grupo de controle, que em função do sorteio, são idênticos em observáveis e não observáveis, *em média, na população.*
- Note que a primeira aleatorização (passo 1) tenta assegurar **validade externa**, enquanto a segunda (passo 2) tenta assegurar **validade interna.**



NÍVEL DA ALEATORIZAÇÃO

- Uma escolha importante é em qual nível se dará a aleatorização: indivíduo, família, cidade, etc.
- Para algumas intervenções não há muita escolha, mas para outras a escolha não é tão clara.



DIFERENTES ESCOLHAS

- Por exemplo, programa de desverminação:
- Dickson and Garner (2000) avaliam programa feito em nível individual dentro das escolas;
- Miguel and Kremer (2004) olham efeito de programa semelhante via phase-in no nível da escola.



FATORES A SEREM CONSIDERADOS PARA A ESCOLHA DO NÍVEL DA ALEATORIZAÇÃO

1. Quanto maior o tamanho dos grupos que são randomizados, maior o tamanho da amostra total que precisa ser alcançado.
 - O nível da aleatorização tem impacto grande no orçamento da avaliação, tornando a aleatorização em nível individual atrativa quando possível.



FATORES A SEREM CONSIDERADOS PARA A ESCOLHA DO NÍVEL DA ALEATORIZAÇÃO

2. Spillovers do grupo de tratamento para o de controle podem viesar a estimação do efeito do tratamento
 - Nesses casos, aleatorização deve ocorrer em nível que evite esses efeitos.



FATORES A SEREM CONSIDERADOS PARA A ESCOLHA DO NÍVEL DA ALEATORIZAÇÃO

- Miguel and Kremer (2004) acharam efeitos muito maiores do que trabalhos anteriores que aleatorizaram em nível do indivíduo.
- Argumentam que como a infestação espalha facilmente entre crianças, o grupo de controle também se beneficiou do tratamento, reduzindo a distância entre controle e tratamento.



FATORES A SEREM CONSIDERADOS PARA A ESCOLHA DO NÍVEL DA ALEATORIZAÇÃO

- Uma outra externalidade que pode ocorrer é que os indivíduos **ino grupo de controle mudam comportamento antecipando que serão tratados no futuro.**
- Pode ser mais fácil evitar isso se a aleatorização for ao nível de município ao invés de indivíduos.



FATORES A SEREM CONSIDERADOS PARA A ESCOLHA DO NÍVEL DA ALEATORIZAÇÃO

3. Aleatorização em nível mais agregado pode ser mais fácil de implementar, mesmo que requeira amostras maiores.
- Ex: Programa é um curso de formação. Se a randomização for ao nível da cidade, todos da cidade fazem o curso junto e cai pela metade o número de lugares em que dou o curso.



DEFININDO O TRATAMENTO

- Parece óbvio, mas não é!
- Você pode oferecer um tratamento, mas nem todos os alocados ao grupo de tratamento aceitarem o mesmo
- Nesse caso, você está estimando o impacto causal da oferta do tratamento (*intention to treat* – ITT)
 - Contorna o problema de no-show e atrito



DEFININDO O TRATAMENTO

- Mas posso estar interessado no efeito de participar de todo o tratamento → impacto efetivo
- Problemas:
 - No-show
 - Atrito
 - Impacto com relação a que? A 2ª melhor opção?
 - Os que saíram ou não apareceram podem ser grupos de controle alternativos. → cautela



AMEAÇAS A ALEATORIZAÇÃO

- Contaminação
 - Se pessoas no tratamento interagirem com as do grupo de controle, há contaminação do grupo de controle, mudando eventualmente os resultados desse grupo.
 - Ex: visitas domiciliares de agente de saúde. Se converso com minha vizinha, repasso as orientações recebidas.



AMEAÇAS A ALEATORIZAÇÃO

- Cross-over: se pessoas passam do grupo de controle ao grupo de tratamento e vice-versa
 - Ex: tamanho de classe: de um ano para o outro, alguns dos alunos inicialmente alocados para classes pequenas vão para classes grandes e vice-versa.
 - Ex: creches no Rio de Janeiro (PB), Liga Solidária. Solução: LATE



AMEAÇAS A ALEATORIZAÇÃO

- Atrito
 - Participantes que decidem abandonar o tratamento são diferentes daqueles que permanecem em características não observáveis que podem potencialmente afetar os resultados.
 - Quanto maior o período de acompanhamento (follow-up), maior o tamanho do atrito.
 - Pode afetar a validade externa, via diminuição do tamanho da amostra, e a validade interna (se a taxa de atrito for muito diferente nos grupos, é indício ruim)



AMEAÇAS A ALEATORIZAÇÃO

- Atrito
 - mesmo que a taxa de atrito seja semelhante nos grupos, o padrão do atrito pode ser distinto
 - se tiver feito uma linha de base antes do tratamento, e as características médias dos que ficaram e saíram no tratamento e controle (4 grupos) for semelhante, sem problemas de validade interna.



EM SUMA

- Para a aleatorização ser válida, a análise precisa focar nos grupos inicialmente criados pela aleatorização.
- Devemos comparar todos os inicialmente alocados ao tratamento com todos inicialmente alocados para o grupo de controle, qualquer que tenha sido seu comportamento e status de tratamento ex-post.



AMEAÇAS A ALEATORIZAÇÃO

- Efeito Hawthorne e John Henry
 - Implicitamente, assumimos que a participação no tratamento per se, não afeta o comportamento e resultados dos participantes
 - Hawthorne: mudança de comportamento dos participantes de um experimento pelo simples fato de serem objeto de estudo
 - John Henry: ocorre quando o grupo de controle muda seu comportamento para “competir” com o do tratamento por estarem insatisfeitos de serem alocados ao grupo de controle.



AMEAÇAS A ALEATORIZAÇÃO

- Efeito Hawthorne e John Henry

Ex: projeto STAR

- Professores sabem que estão sendo monitorados e podem ter feito maior esforço em decorrência disso

Ex: projeto de voucher

- Alunos não contemplados podem se esforçar muito para melhorar seu desempenho e não depender do voucher para conseguir vaga em escola privada.



ANÁLISE EXPERIMENTAL

- Pesquisa experimental em economia se origina da preocupação com a identificação confiável do efeito de programas em face a múltiplos e complexos canais de causalidade.
- Os experimentos possibilitam variar cada um dos fatores por vez e, portanto, fornecem estimativas com alta validade interna do efeito causal.



CROSS-CUTTING DESIGNS

- Uma das inovações institucionais que levaram a crescimento do numero de avaliações aleatorizadas é o uso crescente de um desenho denominado **cross-cutting** (ou **fatorial**).
- No desenho **cross-cutting** diferentes tratamentos são testados simultaneamente com as aleatorizações sendo feitas de sorte que os tratamentos sejam ortogonais entre si.



EXEMPLO

- Programa de desverminação (Miguel and Kremer 2004), que combina duas intervenções: (1) comprimido de vermífugo; (2) conselho para as crianças sobre hábitos de higiene (lavar as mãos, etc).
- Não houve mudança de comportamento nas escolas de tratamento após a intervenção, o que sugere que o componente da intervenção que fez diferença foi o vermífugo mesmo.



Por que isso é importante?

- Estimativas de custo-efetividade mostram que para uma mesma melhoria educacional (correspondente ao aprendizado de 1 ano de escola), há programas custando entre \$3.25 e \$200 por criança. (J-PAL, 2005)
- A intuição pode não ajudar a descobrir qual o melhor tratamento.



Motivação para Experimentos

- Primeiro, adoção de políticas efetivas exige realização de julgamentos acerca da eficácia de componentes individuais dos programas, sem fé excessiva em conhecimento a priori estabelecido.
- Segundo, porém, é difícil aprender sobre tais componentes individuais usando dados observacionais (não experimentais)



Motivação para Experimentos

- Em educação, particularmente, essa dificuldade decorre do fato de que dados observacionais advem de sistemas escolares que adotaram determinado “modelo”, que consiste em mais de um insumo.
- A variação observada nos insumos escolares decorre de tentativas de mudar o modelo , o que em geral, envolve realizar várias mudanças simultâneas.



Exemplo

- Bônus ES → muda bônus e currículo junto, muda critério de seleção de diretores.
- Portanto, dados observacionais não permitem enxergar o efeito de variação dos componentes individuais das políticas.



Implicação

- Assim, dado o alto custo fixo de realização de experimento...
- Vale a pena a realização de experimentos múltiplos que testem variações do programa, mesmo quando você não está interessado no efeito da interação de programas (Miguel Kremer, 2004).



EXEMPLO: SERVIR PARA TESTAR MÚLTIPLAS HIPÓTESES

- ..., com pequeno aumento de custo, pois o principal custo da aleatorização consiste em realizar o baseline e medir as variáveis de resultado.
- Nesse caso, o tamanho da amostra total precisa somente ser grande o suficiente para que haja poder suficiente para identificar o efeito da intervenção de menor efeito.



BANERJEE, DUFLO, COLE, AND LINDEN (2007)

- Testaram em uma mesma amostra (escolas municipais em Vadodara, Índia) o efeito de aulas de recuperação e de ‘Computer Assisted Learning’ sobre desempenho escolar.
- Metade das escolas recebeu ‘remedial education program’ e a outra metade recebeu “computer assisted learning”.



Experimentos como processos dinâmicos



Processo Dinâmico

- Experimento deve ser processo dinâmico, onde primeiro se avalia o programa em linhas gerais.
- Dadas as primeiras respostas, muitas vezes inesperadas, realiza-se uma segunda rodada de análise, para investigar que “partes” do programa funcionam melhor.



Preocupações com Experimento

- Dependência do ambiente
 - Quão generalizáveis são os resultados?
 - Uniformes escolares no Kenia, Mexico e Noruega
 - Validade externa
 - ONGs e governos que aceitam ser parte de experimento são diferentes dos demais (Heckman, 92)
 - Necessidade de replicação em outros contextos. Ex: PROGRESA foi replicado em vários países (Colombia, Nicaragua, Equador e Honduras)



Preocupações com Experimento

- Efeito de Equilíbrio Geral
 - Ex: Voucher escola privada. Se adota em todo país, efeito positivo pode desaparecer, pois satura a rede privada e reduz retorno a educação, via aumento de oferta



Preocupações com Experimento

- Efeito de Equilíbrio Geral
 - Saída: tentar estimar esse efeito de equilíbrio geral
 - Kremer and Muralidharan estudam efeito de voucher com dupla aleatorização: tanto cidades como pessoas na cidade são aleatorizadas
 - Comparando as estimativas dos 2 tratamentos, podem calcular efeito de eq. geral.
 - Porém, pessoas podem se mudar de uma cidade para outra para procurar trabalho



Heterogeneidade do efeito do tratamento

- Efeito médio do tratamento, único parâmetro que se extrai da aleatorização sem nenhuma hipótese adicional, pode não ser o parâmetro de interesse do gestor.



Convencendo o gestor a aleatorizar

- Aspecto ético da aleatorização
- Justiça (igualdade de oportunidades)
- Excesso de demanda (já vai excluir alguém, por algum critério)



Multiplicidade de desenhos

- Loteria com maior probabilidade para os mais vulneráveis
- Aleatorizar regiões
- Abrir inscrição e depois aleatorizar
- Futuros beneficiários como grupos de controle



Multiplicidade de desenhos

- Emparelhar indivíduos em observáveis e aleatorizar quem recebe o tratamento
 - Assegura balanceamento das variáveis
 - Facilita a interpretação dos resultados
 - Reduz risco de perda do desenho experimental (Ex: Glewwe – óculos para cidades chinesas)
 - Aumenta precisão, pois permite estratificação amostral (Imbens et al, On the Benefits of Stratification, 2008)



Multiplicidade de desenhos

- Ex Glewwe
- Em cada condado, cada cidade foi ranqueada de acordo com a renda per capita em 2003. A partir das duas primeiras cidades mais ricas, uma foi aleatoriamente selecionada para o tratamento e a outra para o controle; se repetiu essa análise para todos os pares subsequentes de cidades.
- Em cada cidade, todas as escolas primárias ou eram atribuídas ao grupo de tratamento ou ao grupo de controle.



Inferência em pequenas amostras

- Teste de Permutação de Fisher (1925)
- Fisher propôs o seguinte experimento: preparem 8 xícaras de chá com leite, sendo quatro delas em que se coloca primeiro o chá e nas outras 4 primeiro o leite.
- Um observador deverá receber as 8 xícaras de chá aleatoriamente e separá-las nos dois grupos: as que receberam chá primeiro e as demais



Inferência em pequenas amostras

- Hipótese nula: O observador não tem essa capacidade.
- Se H_0 for válida, ele vai escolher aleatoriamente quatro xícaras entre as oito disponíveis. Isso pode ser feito de 70 jeitos diferentes. (Arranjo de 8, 4 a 4)
- Quando H_0 é válida, vai escolher as 4 xícaras corretamente com probabilidade $1/70$.



Teste de Permutação de Fisher

- Definição: Uma estatística T é uma função conhecida de $T(W, Y_{\text{obs}}, X)$ de tratamentos atribuídos (W), resultados observados (Y_{obs}) e variáveis pré-tratamento (X).
- A ideia é comparar o valor da estatística T para a atribuição ao tratamento efetivamente realizada com a distribuição de todas as possíveis atribuições ao tratamento.



Teste de Permutação de Fisher

- Isso possibilita calcular a probabilidade de se observar um valor da estatística T tão extremo (ou mais) que o efetivamente observado.
-
- Ou seja, permite calcular o p-valor para a hipótese nula de que o tratamento específico não possui efeito.



Teste de Permutação de Fisher

- O motivo para os resultados Y_{obs} e T serem aleatórios é que há um mecanismo aleatório de atribuição do tratamento que determina qual dos dois resultados potenciais vai ser revelado para cada indivíduo.
- Uma vez que todos os resultados potenciais são conhecidos se a hipótese nula for verdadeira, não é necessário fazer hipóteses sobre a distribuição dos resultados.



Teste de Permutação de Fisher

- Uma vez que o experimento foi realizado o pesquisador precisa somente fazer 3 escolhas:
 - A) Hipótese nula a ser testada
 - B) a estatística a ser utilizada
 - C) Definir o que “pelo menos tão extremo quanto”



Teste de Permutação de Fisher

4.1 A SIMPLE EXAMPLE WITH TWO UNITS

Let us consider a simple example with 2 units in a completely randomized experiment. Suppose that the first unit got assigned $W_1 = 1$ and so we observed $Y_1(1) = y_1$. The second unit therefore was assigned $W_2 = 1 - W_1 = 0$ and we observed $Y_2(0) = y_2$. We are interested in the null hypothesis that there is no effect whatsoever of the treatment. Under that hypothesis the two unobserved potential outcomes become known for each: $Y_1(0) = Y_1(1) = y_1$ and $Y_2(1) = Y_2(0) = y_2$. Now consider a statistic. The most obvious choice is difference between the observed value under treatment and the observed value under control:

$$T = W_1 \cdot (Y_1(1) - Y_2(0)) + (1 - W_1) \cdot (Y_2(1) - Y_1(0)),$$



Teste de Permutação de Fisher

- O valor da estatística dado a atribuição observada ao tratamento ($W_1 = 1$) é:
- $T_{\text{obs}} = Y_1(1) - Y_2(0) = y_1 - y_2$.
- Agora considere o conjunto de todas as possíveis atribuições ao tratamento
- Nesse caso, só há duas possibilidades para $\mathbf{W} = (W_1, W_2)$.



Teste de Permutação de Fisher

- A primeira é o vetor de atribuição ao tratamento efetivamente ocorrido $\mathbf{W}_1 = (1, 0)$;
- A outra é o contrário: $\mathbf{W}_2 = (0, 1)$;
- Para cada um desses vetores, podemos calcular *qual seria o valor da estatística*:
- Caso 1: $T_{\mathbf{W}_1} = y_1 - y_2$
- Caso 2: $T_{\mathbf{W}_2} = y_2 - y_1$



Teste de Permutação de Fisher

- Ambos os vetores de atribuição possuem a mesma probabilidade (50%). Assim, a distribuição da estatística T , mantida a hipótese nula é dada por:

$$Pr(T = t) = \begin{cases} 1/2 & \text{if } t = y_1 - y_2, \text{ or } t = y_2 - y_1, \\ 0 & \text{otherwise.} \end{cases}$$

- Sob H_0 , conhecemos a distribuição inteira da estatística T . Nesse caso, o p-valor é 1/2



Teste de Permutação de Fisher

- Assim, o resultado não parece extremo. Sempre com $n=2$, não importa H_0 , o resultado observado nunca vai parecer “extremo”.
- Suponha agora que temos um experimento aleatorizado com $2N$ indivíduos, N recebendo o tratamento e N no grupo de controle.



Teste de Permutação de Fisher

- Existem $\binom{2N}{N}$ modos diferentes de atribuir o tratamento aos indivíduos [ou seja há $\binom{2N}{N}$ vetores de atribuição], quando no exemplo anterior havia apenas 2.
- Agora o p-valor pode ser tão pequeno quanto $1/\binom{2N}{N}$



Teste de Permutação de Fisher

- Ex: $N=6$ Tabela a seguir possui observações de uma aleatorização para avaliar o efeito de programa educativo de TV na habilidade de leitura dos alunos. Os dados são por classe da escola
- Resultado de interesse: nota média de leitura da turma; 50% das turmas receberam o tratamento.



Teste de Permutação de Fisher

- Características medidas:
 - Cidade da turma: Fresno, OH ou Youngstown, OH
 - Série da turma
 - Notas de leitura antes e depois do experimento

Unit	Treatment	Fresno/Youngstown	Pre-test Score	Post-test Score
------	-----------	-------------------	----------------	-----------------

1	0	F	12.3	55.0
2	1	F	16.5	70.0
3	0	F	18.7	72.0
4	1	F	51.4	66.0
5	0	F	18.7	72.7
6	1	F	19.4	78.9

Teste de Permutação de Fisher

- H_0 : programa não tem efeito em nenhuma turma i
 $\rightarrow Y_i(0) = Y_i(1)$ para todo $i = 1, \dots, 6$.
- Sob H_0 :

Unit	Potential Outcomes		Actual Treatment	Observed Outcome
	$Y_i(0)$	$Y_i(1)$		
1	55.0	(55.0)	0	55.0
2	(70.0)	70.0	1	70.0
3	72.0	(72.0)	0	72.0
4	(66.0)	66.0	1	66.0
5	72.7	(72.7)	0	72.7
6	(78.9)	78.9	1	78.9



Teste de Permutação de Fisher

- Vamos testar H_0 usando a diferença de resultados entre tratados e controles como estatística:

$$T_1 = \sum_{i=1}^6 W_i \cdot Y_i^{\text{obs}} / 3 - \sum_{i=1}^6 (1 - W_i) \cdot Y_i^{\text{obs}} / 3 = \bar{y}_1 - \bar{y}_0,$$

where

$$\bar{y}_w = \sum_{i=1}^N 1\{W_i = w\} \cdot Y_i^{\text{obs}} / \sum_{i=1}^N 1\{W_i = w\}.$$



Teste de Permutação de Fisher

- Sob H_0 , podemos calcular o valor da estatística sob cada vetor de atribuição do tratamento \mathbf{W} . Há $\binom{6}{3}=20$ possibilidades de atribuição. A tabela a seguir lista todas as possibilidades de 3 classes recebendo tratamento e três classes no controle
- Para cada caso, calculamos o valor da estatística



Teste de Permutação de Fisher

W_1	W_2	W_3	W_4	W_5	W_6	$(\sum W_i \cdot Y_i^{obs} - \sum (1 - W_i) \cdot Y_i^{obs})/3$
0	0	0	1	1	1	5.1
0	0	1	0	1	1	6.9
0	0	1	1	0	1	9.5
0	0	1	1	1	0	0.9
0	1	0	0	1	1	6.4
0	1	0	1	0	1	9.1
0	1	0	1	1	0	0.5
0	1	1	0	0	1	10.9
0	1	1	0	1	0	2.3
0	1	1	1	0	0	4.9
1	0	0	0	1	1	-4.9
1	0	0	1	0	1	-2.3
1	0	0	1	1	0	-10.9
1	0	1	0	0	1	-0.5
1	0	1	0	1	0	-9.1
1	0	1	1	0	0	-6.4
1	1	0	0	0	1	-0.9
1	1	0	0	1	0	-9.5
1	1	0	1	0	0	-6.9
1	1	1	0	0	0	-5.1



Teste de Permutação de Fisher

- Primeira linha é o efetivamente observado: nota do tratamento é 5.1 pontos maior do que a do controle
- O quão extremo é esse valor, de tal sorte que deveríamos observar ganhos tão grandes quanto esse, caso a atribuição do tratamento fosse diferente, sob H_0 (ou seja, se não houver efeito do tratamento)?



Teste de Permutação de Fisher

- Há 6 vetores dos 20 que gerariam diferenças maiores do que as efetivamente observadas entre tratados e controles
- $P\text{-valor} = 6/20 = 0.30$
- Portanto, não rejeitamos a hipótese nula!

