EPI13 - Seminários de Doutorado III

Estrutura

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística e Programa de Pós-Graduação em Epidemiologia

Vitória da Conquista, 2019

Uma População

Uma população

- Os $885 = 439 + 446$ pacientes no Estudo ProCESS são uma **população pequena** e **finita**, não muito diferente da população de pessoas que vivem em uma cidade pequena.
- Poderíamos nos referir aos pacientes pelo nome, mas é mais conveniente numerá-los: $i = 1, 2, \ldots, I = 885$.
  - Aqui, i faz referência a um indivíduo, e I se refere ao número total de indivíduos, $I = 885$.
- Nos exemplos deste capítulo, o indivíduo $i = 17$ é o nosso amigo Harry!
- Ao substituir 885 pacientes por I pacientes, podemos ser fiéis ao descrever o estudo ProCESS, reconhecendo ao mesmo tempo que muitos detalhes do estudo ProCESS são incidentais - por exemplo, o tamanho da amostra - e que o que estamos dizendo é tão verdadeiro quanto qualquer experimento formado pelo lançamento de moedas honestas.

Covariáveis

Covariáveis

- x_i representará as covariáveis observadas para o paciente i.
- Na Tabela 1.1, existem nove covariáveis para cada paciente i, e x_i registra os valores dessas nove covariáveis para o paciente i.
- Várias covariáveis são atributos que podem estar **presentes** ou **ausentes** em vez de números, mas é costume registrar um atributo como 1 se estiver presente e 0 se estiver ausente.
  - A média de um atributo é a proporção de vezes que o atributo está presente (e 100 vezes essa média é a porcentagem), como na Tabela 1.1.

Covariáveis

Relembrando: Tabela 1.1

Table 1.1. Covariate balance in the ProCESS Trial

Treatment group	*Aggressive*	*Less aggressive*
Sample size (number of patients)	439	446
Age (years, mean)	60	61
Male (%)	53	56
Came from nursing home (%)	15	16
Sepsis source		
Pneumonia (%)	32	34
Urinary tract infection (%)	23	20
Intra-abdominal infection (%)	16	13
APACHE II score (mean)	20.8	20.6
Systolic blood pressure (mean)	100	102
Serum lactate (mmol / liter, mean)	4.8	5

Covariáveis

Table 2.1. The value of the nine observed covariates x_{17} for patient 17

	Background			*Source of sepsis*			*Physiology*		
Patient	*Age*	*Male*	*Nursing home*	*Pn*	*UTI*	*A*	*APACHE II*	*Systolic BP*	*Serum lactate*
x_{17}	52	1	0	1	0	0	23	96.1	5.3

A = intra-abdominal infection; APACHE II = Acute Physiology and Chronic Health Evaluation II; BP = blood pressure; Pn = pneumonia; UTI = urinary tract infection.

- Cada paciente i dos pacientes $I = 885$ tem tal tabela x_i de nove números descrevendo o paciente i.

Covariáveis não medidas

- u_i representará as **covariáveis não observadas** para o paciente i.
- A estrutura de u_i é semelhante à estrutura de x_i, mas u_i é uma covariável que não medimos.
- O que há em u_i? Talvez ...
  - ... u_i inclua um indicador, 1 ou 0, de uma variante de um gene relevante para sobreviver ao choque séptico, talvez um gene cuja importância **ainda não foi descoberta**;
  - ... u_i indique o tipo específico de bactéria responsável pela infecção, incluindo sua resistência a vários antibióticos;
  - ... u_i registre a extensão da experiência do médico residente envolvido no cuidado do paciente i;
  - ... u_i descreva o suporte social disponível para o paciente i.

Covariáveis: comentários

1. As covariáveis, x_i ou u_i, existem em uma única versão.
   - Em particular, o paciente i $= 17$ teria o x_i dado na Tabela 2.1 se Harry é aleatorizado para o tratamento agressivo ou para o tratamento menos agressivo.
2. Em um estudo completamente aleatorizado como o estudo ProCESS, a chance de qualquer paciente receber o tratamento agressivo **é a mesma** que a chance de esse paciente receber o tratamento menos agressivo.
   - Essa chance **não depende** de x_i e nem de u_i.
   - Sabemos disso porque atribuímos tratamentos lançando uma moeda honesta.

- A chance de que Harry seja designado para tratamento agressivo é $1/2$, e **não importa**, no que diz respeito a essa chance, que Harry tem 52 anos com uma escore APACHE II de 23.
  - As coisas seriam diferentes na ausência de atribuição de tratamentos, mas o ensaio ProCESS foi aleatorizado.

Atribuições de Tratamento

Atribuições de tratamento

- Z_i registra o tratamento atribuído ao paciente i.

$$Z_i = \begin{cases} 1, & \text{se o paciente } i \text{ foi atribuído ao tratamento agressivo,} \\ 0, & \text{se o paciente } i \text{ foi atribuído ao tratamento menos agressivo.} \end{cases}$$

- Harry ($i = 17$) foi atribuído ao tratamento agressivo, $Z_{17} = 1$.

Atribuições de tratamento

- Podemos reformular Z_i em termos mais genéricos

$$Z_i = \begin{cases} 1, & \text{se o paciente } i \text{ foi atribuído ao tratamento,} \\ 0, & \text{se o paciente } i \text{ foi atribuído ao controle.} \end{cases}$$

- m representará o número de pacientes no grupo tratado (no estudo ProCESS, $m = 439$).

Atribuições de tratamento

- O estudo ProCESS atribuiu tratamentos aleatoriamente lançando uma moeda honesta, de modo que Z_i fosse uma quantidade aleatória assumindo o valor $Z_i = 1$ com probabilidade $1/2$ e o valor $Z_i = 0$ com probabilidade $1/2$.
  - $\Pr(Z_i = 1) = 1/2 = \Pr(Z_i = 0)$.
- Utilizaremos $\pi_i = \Pr(Z_i = 1)$ para designar a probabilidade do paciente i ser designado ao grupo tratado.
  - Como o estudo ProCESS é um ensaio completamente aleatorizado, $\pi_i = 1/2$ para $i = 1, \ldots, I$ em que $I = 885$.
- Grande parte da complexidade da inferência causal surge quando π_i varia de pessoa para pessoa de maneiras que não compreendemos completamente.

Atribuições de tratamento

- Uma tarefa é usar todos os aspectos do passado dos indivíduos para criar grupos tratados e de controle idênticos, o que não pode ser feito.
- A segunda tarefa é garantir que absolutamente nenhum aspecto do passado dos indivíduos influencie sua designação de tratamento, o que é simples: lançamos uma moeda honesta.
- Felizmente, como será visto em discussão futura, o sucesso na segunda tarefa é tudo o que é necessário para a inferência causal.

Efeitos Causados por Tratamentos

Desfechos potenciais sob tratamentos alternativos

- **Relembrando:** efeitos causais são expressos como comparações de desfechos (respostas) potenciais sob tratamentos concorrentes.
- r_{C_i} e r_{T_i} representam os desfechos potenciais no indivíduo i.
  - r_{C_i} é a desfecho que teria sido observado caso o paciente i fosse atribuído ao protocolo menos agressivo (controle).
  - r_{T_i} é a desfecho que teria sido observado caso o paciente i fosse atribuído ao protocolo agressivo (tratamento).
- Se o paciente i foi atribuído ao tratamento, então irá exibir o desfecho r_{T_i}, mas não exibirá r_{C_i}.
- Se o paciente i foi atribuído ao controle, então irá exibir o desfecho r_{C_i}, mas não exibirá r_{T_i}.

Desfechos potenciais sob tratamentos alternativos

- No estudo ProCESS,
  - $r_{T_i} = 1$, se o paciente i morreu no hospital antes de 60 dias se o paciente tivesse recebido o protocolo agressivo;
  - $r_{T_i} = 0$, se o paciente i recebeu alta do hospital antes de 60 dias ou sobreviveu no hospital por 60 dias se o paciente tivesse recebido o protocolo agressivo;
  - $r_{C_i} = 1$, se o paciente i morreu no hospital antes de 60 dias se o paciente tivesse recebido o protocolo menos agressivo;
  - $r_{C_i} = 0$, se o paciente i recebeu alta do hospital antes de 60 dias ou sobreviveu no hospital por 60 dias se o paciente tivesse recebido o protocolo menos agressivo.

Desfechos potenciais sob tratamentos alternativos

- Para Harry, paciente $i = 17$, nós vemos $r_{T_{17}}$ mas não vemos $r_{C_{17}}$, pois Harry foi designado ao protocolo agressivo, $Z_{17} = 1$.
- Desfechos potenciais são também chamados de **contrafactuais** porque descrevem o que aconteceu ao Harry se, **contrário ao fato**, Harry tivesse recebido o tratamento que ele não recebeu. (!)

Desfechos potenciais sob tratamentos alternativos

- $(r_{C_i}, r_{T_i}) = (0, 1)$, então diríamos que o tratamento menos agressivo **causou** a morte do paciente i no hospital (antes de 60 dias).
  - Efeito positivo para o paciente i.
- $(r_{C_i}, r_{T_i}) = (1, 0)$, então diríamos que o tratamento agressivo **causou** a morte do paciente i no hospital (antes de 60 dias).
  - Efeito negativo para o paciente i.
- Se cada paciente i **apresenta** $r_{T_i} = r_{C_i} = 1$ ou $r_{T_i} = r_{C_i} = 0$, então não importa o tratamento recebido.
  - "Dois caminhos sempre terminariam no mesmo lugar".

Desfechos potenciais sob tratamentos alternativos

"Dois caminhos sempre terminariam no mesmo lugar"

- A **hipótese de Fisher** de nenhuma diferença nos efeitos dos dois tratamentos.
- A hipótese de Fisher afirma que $r_{T_i} = r_{C_i}$ para cada paciente i no estudo ProCESS.
- Nós escrevemos esta hipótese como $H_0 : r_{T_i} = r_{C_i}, i = 1, \ldots, I$, em que $I = 885$ no estudo ProCESS.

Desfechos potenciais sob tratamentos alternativos

- O par de desfechos potenciais (r_{C_i}, r_{T_i}) será descrito como o **efeito causal** (individual).
- Chamaremos $\delta_i = r_{T_i} - r_{C_i}$ de a **diferença do efeito causal**.
  - A hipótese de Fisher é equivalente a $\delta_i = 0$ para todo i.
- O tratamento agressivo salvou o paciente i se $\delta_i = r_{T_i} - r_{C_i} = 0 - 1 = -1$.
- O tratamento agressivo causou a morte do paciente i se $\delta_i = r_{T_i} - r_{C_i} = 1 - 0 = 1$.
- Note que vemos uma parte de (r_{C_i}, r_{T_i}), r_{C_i} ou r_{T_i}, mas nunca vemos $\delta_i = r_{T_i} - r_{C_i}$.

Desfechos potenciais sob tratamentos alternativos

- Para cada paciente, nós vemos apenas um dos dois desfechos potenciais.
  - Se o paciente i recebeu o tratamento agressivo, vemos r_{T_i}, mas não o r_{C_i}.
  - Caso o paciente i tivesse recebido o tratamento menos agressivo, veríamos r_{C_i}, mas não veríamos r_{T_i}.
- No estudo ProCESS vemos r_{T_i} para 439 pacientes, e vemos r_{C_i} para os demais 446 pacientes.
- Nós nunca vemos ambos r_{C_i} e r_{T_i} para o mesmo paciente.
- Como discutido anteriormente, é isso que torna a inferência causal difícil: declarações causais referem-se a r_{T_i} e r_{C_i} conjuntamente, mas nunca vemos r_{T_i} e r_{C_i} conjuntamente[1].

[1] **Problema fundamental da inferência causal** (ver Holland, P. Statistics and Causal Inference. *JASA*, 81:945-960, 1986)

O desfecho observado

- R_i será utilizado para denotar o desfecho observado do paciente i.

$$R_i = Z_i \times r_{T_i} + (1 - Z_i) \times r_{C_i}.$$

- Esta relação é conhecida na literatura como a **suposição de consistência**.
- O que observamos são os Z_i e R_i de cada participante.

Table 1.2. In-hospital mortality outcomes in the ProCESS Trial

	In-hospital 60-day mortality			
Treatment group	*In-hospital death*	*Other*	*Total*	*Death rate (%)*
Aggressive	92	347	439	21.0
Less aggressive	81	365	446	18.2
Total	173	712	885	

Médias em Populações e Amostras

Tamanho da população I e tamanhos das amostras m e $I - m$

- O número de pacientes no grupo tratado m pode ser expresso em função dos Z_i

$$m = \sum_{i=1}^{I} Z_i = Z_1 + Z_2 + \ldots + Z_I.$$

- De forma análoga, temos o número de pacientes no grupo controle $I - m$

$$I - m = \sum_{i=1}^{I} (1 - Z_i) = (1 - Z_1) + (1 - Z_2) + \ldots + (1 - Z_I).$$

- A proporção de pessoas designadas ao grupo tratamento é $m/I = (1/I) \sum_{i=1}^{I} Z_i$.

Médias populacionais ($\bar{v}$)

- Suponha que temos uma variável v_i que está definida para cada pessoa na população, $i = 1, \ldots, I$, em que $I = 885$ no estudo ProCESS.
- Por exemplo, o escore APACHE II foi medido para todos os $I = 885$ pacientes na Tabela 1.1.
- A média dos $I = 885$ escores APACHE II é $20,7$, e geralmente escrevemos $\bar{v}$ para uma média.
- A média, $\bar{v}$, é a soma do v_i dividido pelo número de termos na soma (I).
    - Ou seja, $\bar{v} = 20,7 = (1/885)(v_1 + v_2 + \ldots + v_{885}) = (1/I)(v_1 + v_2 + \ldots v_I)$.

Médias amostrais ($\hat{v}_T$ e $\hat{v}_C$)

- A Tabela 1.1 fornece o escore APACHE II médio em cada um dos dois grupos de tratamento, o grupo de tratamento agressivo com $Z_i = 1$ e o grupo de tratamento menos agressivo com $Z_i = 0$.
- Cada um deles é uma **média amostral** para uma **amostra aleatória simples** do população finita de $I = 885$ pacientes no estudo como um todo.
- Escrevemos $\hat{v}_T$ para a média dos valores $m = 439$ v_i no grupo tratado, o v_i para os $m = 439$ pessoas com $Z_i = 1$.
  - Na Tabela 1.1, $\hat{v}_T = 20,8$ para o escore APACHE II médio no grupo tratado.
  - De forma geral, temos $\hat{v}_T = (1/m)\sum_{i=1}^{I} Z_i v_i$.
- Da mesma forma, temos Escrevemos $\hat{v}_C$ para a média dos valores $I - m = 446$ v_i no grupo controle ($Z_i = 0$).
  - Na Tabela 1.1, $\hat{v}_C = 20,6$ para o escore APACHE II médio no grupo controle.
  - De forma geral, temos $\hat{v}_C = \{1/(I - m)\}\sum_{i=1}^{I} (1 - Z_i) v_i$.

Equilíbrio na distribuição das covariáveis em experimentos aleatorizados

- Temos três médias dos escores APACHE II:
  - a média populacional, $\bar{v} = 20,7$
  - a média no grupo aleatoriamente designada para o grupo de tratamento, $\hat{v}_T = 20,8$
  - e a média no grupo aleatoriamente designado para controle, $\hat{v}_C = 20,6$.
- As três médias são similares porque os grupos tratados e controle foram formados de forma aleatória, lançando moedas honestas.

Equilíbrio na distribuição das covariáveis em experimentos aleatorizados

- Anteriormente, interpretamos esse cálculo como uma uma confirmação do equilíbrio da covariável que esperávamos que a designação aleatória produzisse.
  - Esperávamos que os dois grupos fossem semelhantes porque foram formados aleatoriamente e vimos que eles eram semelhantes (Fórum de discussão no Moodle).
- Quando discutimos **covariáveis não observadas**, esperávamos que os dois grupos fossem semelhantes porque foram formados de forma aleatória, mas não pudemos mais confirmar nossa expectativa por meio de inspeção direta.

Equilíbrio na distribuição das covariáveis em experimentos aleatorizados

- A relação entre $\hat{v}_T$, $\hat{v}_C$ e $\bar{v}$ tem um terceiro uso.
- Se pudéssemos ver os escores APACHE II em apenas um dos dois grupos de tratamento, então poderíamos usar a média nesse grupo para estimar a média populacional;
  - Poderíamos usar $\hat{v}_T = 20,8$ ou $\hat{v}_C = 20,6$ para estimar $\bar{v} = 20,7$.
- Este terceiro uso não é importante para os escores do APACHE II porque temos todos eles, mas é importante na inferência causal.

Estimando médias populacionais de médias de amostras aleatórias

- Escores APACHE II "esquecidos", amostras aleatórias e estimativa da média populacional pela média amostral.
  - Da teoria da amostragem de populações finitas, sob certas condições, a média de uma grande amostra aleatória é uma boa estimativa de uma média populacional.
- A teoria da amostragem depende fortemente do uso de amostragem aleatória (o uso do lançamento da moeda).
  - se tivéssemos $m = 439$ medidas v_i de uma população de $I = 885$ medições, mas **não** tivéssemos uma amostra aleatória (escolhida por lançamentos de moeda), então não teríamos razão para acreditar que a média da amostra é próxima da média populacional.

Estimando médias populacionais de médias de amostras aleatórias

- Essa pequena história estranha sobre as escores APACHE II esquecidos acaba sendo quase idêntica à maneira como estimamos os **efeitos causais médios** para os pacientes com $\mathbf{I = 885}$ quando não podemos ver o efeito causal de nenhum paciente.
- Vemos os resultados de sobrevida sob tratamento agressivo r_{T_i} apenas para os $m = 439$ pacientes que receberam tratamento agressivo, e vemos resultados de sobrevida sob tratamento menos agressivo apenas para os $I - m = 446$ pacientes que receberam tratamento menos agressivo,
  - Porque cada grupo é uma amostra aleatória da população de $I = 885$ pacientes, podemos usar as duas médias amostrais para estimar uma média populacional.

Estimando médias populacionais de médias de amostras aleatórias

- Tal como acontece com os escores APACHE II, o elemento-chave é a **atribuição aleatória de tratamentos**, que produz duas amostras aleatórias da população de $I = 885$ pessoas.
- Ao contrário dos escores do APACHE II, na inferência causal não temos todos os v_i, mas temos as duas médias amostrais das duas amostras aleatórias complementares.

Efeitos Causais Médios

Qual teria sido a taxa de mortalidade se todos os pacientes tivessem recebido o tratamento agressivo?

- Duas quantidades causais de interesse que não conseguimos calcular dos dados observados:
  - A taxa de mortalidade de todos os $I = 885$ pacientes caso eles tivessem sido designados ao protocolo agressivo, $\bar{r}_T = (1/I)\sum_{i=1}^{I} r_{T_i}$;
  - A taxa de mortalidade de todos os $I = 885$ pacientes caso eles tivessem sido designados ao protocolo menos agressivo, $\bar{r}_C = (1/I)\sum_{i=1}^{I} r_{C_i}$.
- No entanto, podemos calcular as taxas de mortalidade amostrais, $\hat{r}_T = (1/m)\sum_{i=1}^{I} Z_i r_{T_i}$ (0,21 no estudo ProCESS), e $\hat{r}_C = \{1/(1-m)\}\sum_{i=1}^{I} (1-Z_i) r_{C_i}$ (0,182).

Qual teria sido a taxa de mortalidade se todos os pacientes tivessem recebido o tratamento agressivo?

- Temos razões para esperar que a quantidade que podemos calcular, $\hat{r}_T = 0,210$, seja uma boa estimativa da quantidade que queremos mas não podemos calcular, $\bar{r}_T$, porque $\hat{r}_T$ é a média de uma amostra aleatória de $m = 439$ dos r_{T_i} de uma população composta por todos os $I = 885$ r_{T_i}.
- E pelas mesmas razões, para acreditamos que $\hat{r}_C = 0,182$ é uma boa estimativa para $\bar{r}_C$.

Qual é a diferença média na mortalidade causada pela diferença nos tratamentos?

- Se $\delta_i = r_{T_i} - r_{C_i}$ e $\overline{\delta} = (1/I)\sum_{i=1}^{I} \delta_i$ é a média da diferença de efeito causal.
- Como podemos estimar $\overline{\delta}$ se nunca vemos um único $\delta_i = r_{T_i} - r_{C_i}$?
- Note que

$$\begin{aligned}
\overline{\delta} = \frac{1}{I}\sum_{i=1}^{I} \delta_i &= \frac{1}{I}\sum_{i=1}^{I} r_{T_i} - r_{C_i} \\
&= \frac{1}{I}\sum_{i=1}^{I} r_{T_i} - \frac{1}{I}\sum_{i=1}^{I} r_{C_i} \\
&= \overline{r}_T - \overline{r}_C.
\end{aligned}$$

- A quantidade $\overline{\delta} = \overline{r}_T - \overline{r}_C$ é chamado de **efeito médio do tratamento**.

Qual é a diferença média na mortalidade causada pela diferença nos tratamentos?

- Na Tabela 1.2, estimamos $\bar{r}_T$ por $\hat{r}_T = 21,0\%$ e estimamos $\bar{r}_C$ por $\hat{r}_C = 18,2\%$, então estimamos $\bar{\delta} = \bar{r}_T - \bar{r}_C$ em $21,0\% - 18,2\% = 2,8\%$.
- A estimativa pontual sugere que o tratamento agressivo aumentou a taxa de mortalidade intra-hospitalar em 2,8% sobre o que teria sido com o tratamento menos agressivo.
  - $\hat{r}_T$ é apenas uma estimativa de $\bar{r}_T$ e as estimativas são um pouco erradas.
  - $\hat{r}_C$ é apenas uma estimativa de $\bar{r}_C$, então $\hat{r}_C$ também é um pouco errada.

Qual é a diferença média na mortalidade causada pela diferença nos tratamentos?

- Então, ainda temos que nos perguntar se a nossa **estimativa** $\hat{r}_T - \hat{r}_C = 2,8\%$ do efeito médio do tratamento $\bar{\delta}$ poderia realmente estimar um **efeito populacional** igual a 0.
  - $\hat{r}_T - \hat{r}_C = 2,8\%$ é compatível com $\bar{\delta} = 0$?
  - $\hat{r}_T - \hat{r}_C = 2,8\%$ é compatível com a hipótese de Fisher de que $\delta_i = 0$ para $i = 1, \ldots, I$?
- Precisamos nos perguntar se essa diferença, 2,8%, poderia ser devida ao acaso - devido aos lançamentos de moeda que dividiram a população de $I = 885$ pacientes em duas amostras aleatórias de tamanho $m = 439$ e $I - m = 446$ pacientes.
- Será que $0 = \bar{\delta} = \bar{r}_T - \bar{r}_C$ mas $\hat{r}_T - \hat{r}_C = 2,8\%$ por causa de uma sequência infeliz de lançamentos de moeda, Z_i, na atribuição de tratamentos?
  - Cenas do próximo capítulo!

Por enquanto é só!