Métodos Quantitativos

Aula 08. Testes de hipóteses

Pedro H. G. Ferreira de Souza

pedro.ferreira@ipea.gov.br

Mestrado Profissional em Políticas Públicas e Desenvolvimento

Instituto de Pesquisa Econômica Aplicada (Ipea)

14 nov. 2022

Recapitulação

Introdução

Testes de hipóteses para médias

Mais propriedades importantes

Testes de significância para proporções

Tipos de erros em testes

Comparações entre médias

Limitações de testes de significância

Próxima aula

Recapitulação

Introdução

Testes de hipóteses para médias

Mais propriedades importantes

Testes de significância para proporções

Tipos de erros em testes

Comparações entre médias

Limitações de testes de significância

Próxima aula

Aula passada

Intervalos de confiança

- Proporções
  - Aproximação normal
  - Score de Wilson com correções
  - IC para proporções multinomiais
- Médias
  - Teste t com $n - 1$ graus de liberdade
  - IC com $\bar{y} \pm t \cdot se = t \cdot \frac{s}{\sqrt{n}}$

Testes de hipóteses

Aula 1

Perguntas de pesquisa $\rightarrow$ Teoria $\rightarrow$ Pesquisa empírica $\rightarrow$ Conclusões

- Uma hipótese é uma afirmação - descritiva ou explicativa - teoricamente embasada sobre uma relação que esperamos observar
- Uma hipótese precisa ser **falsificável** $\rightarrow$ os dados que observamos são compatíveis com nossas previsões?

Aula 07

Em um mundo probabilístico em que só analisamos um de muitas amostras possíveis, há uma **incerteza** inerente aos nossos dados.

- ICs quantificam o grau de incerteza amostral dos resultados

Testes de significância

Formalização de **testes de hipóteses**: comparamos nossas estimativas de ponto com valores preditos pelas nossas hipóteses.

- Não podemos provar que nossa hipótese é **verdadeira**, por isso o objetivo dos testes em geral é tentar **falsificá-la**.
- Tipicamente, perguntamos algo como: "dada nossa amostra, qual a probabilidade que nosso resultado X seja só um acaso?"

Exemplo

Suponha um estudo sobre diferenciais salariais por gênero que estima um prêmio salarial de X% para homens. O teste para tentar falsificar essa hipótese vai ser algo como: "qual a probabilidade de encontrarmos esse resultado se no mundo real o parâmetro for igual a zero?"

As cinco partes de um teste de hipóteses

1. Pressupostos

- Tipo de dados; aleatorização; distribuição populacional; amostra

2. Hipóteses

- Hipótese nula e hipótese alternativa; “prova por contradição”

3. Estatística de teste

- Distância entre nossa estimativa e o parâmetro sugerido por H_0, normalmente medido em múltiplos do erro padrão

4. Cálculo do p-valor

- Probabilidade de obtermos nossa estimativa (ou valor mais extremo) se a hipótese nula for verdadeira

5. Conclusão

Etapas 1 e 2

1. Presupposstos

- Amostra aleatória
- Variável quantitativa com distribuição normal na população

2. Hipóteses

- Hipótese nula H_0 de que a média populacional μ tem a forma $H_0 : \mu = \mu_0$, em que μ_0 é um valor qualquer
- Hipótese alternativa de $H_0 : \mu \neq \mu_0$ (teste de duas caudas)

Etapa 3

3. Estatística de teste

- A **estimativa de ponto** é a média amostral $\bar{y}$
- A **distribuição amostral** de $\bar{y}$ é $N \sim (\mu, \frac{\sigma^2}{n})$ se y tiver distribuição normal na população e aproximadamente normal se y tiver outra distribuição, mas a amostra for relativamente grande (TCL)
- Se H_0 for verdadeira, é pouco provável que nossa estimativa $\bar{y}$ esteja muitos erros padrão distante de μ_0
- Logo, as evidências sobre H_0 são avaliadas por um **teste t**:

$$t = \frac{\bar{y} - \mu_0}{se} \quad \text{com} \quad se = \frac{s}{\sqrt{n}} \quad \text{e} \quad df = n - 1$$

Etapa 4

4. Cálculo do p-valor

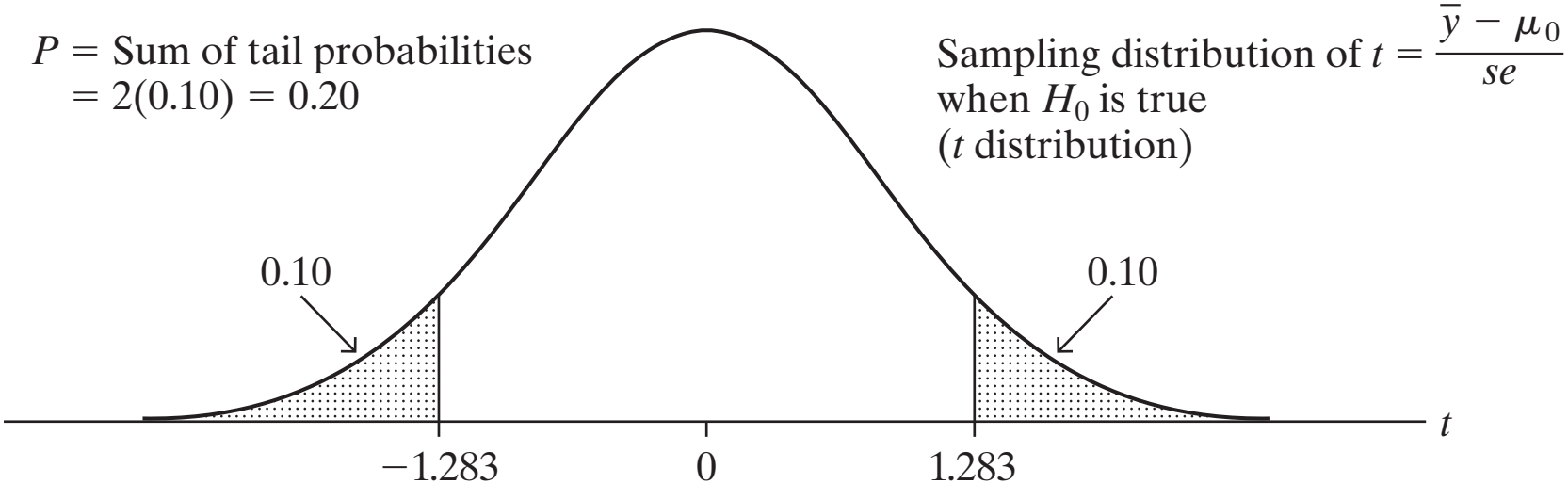
- Calculamos o p-valor supondo que H_0 é verdadeira: qual a probabilidade de que a estatística de teste seja igual (ou mais extrema) que t nesse caso?
- Para $H_0 : \mu \neq \mu_0$, temos um teste em duas caudas, pois valores extremos ocorrem se $\bar{y}$ for muito maior ou muito menor que μ_0
- O **t-score** mais compatível com H_0 é $t = \frac{\bar{y}-\mu_0}{se} = 0$. Nesse caso, o p-valor é $p = 1$.

Exemplo em Agresti, p. 144

Suponha $t = \frac{\bar{y}-\mu_0}{se} = 1.283$, com $n = 369$. O p-valor é a probabilidade de $t \geq 1.283$ ou $t \leq -1.283$, isto é, $|t| \geq 1.283$), com $df = 368$. ou seja, $p = 0.20$.

Etapa 4

4. Cálculo do p-valor




```
# Codigo para R
pt(q = - 1.283, df = 368) + (1 - pt(q = 1.283, df = 368))
```

```
## [1] 0.2002995
```

Etapa 5

5. Conclusão

- Quando menor o p-valor, mais forte é a evidência **contra** H_0 e a favor de H_a.
- A interpretação do p-valor depende da escolha de um **nível de significância**. Normalmente, escolhemos $1 - .95 = .05$.
  - Ou seja, **rejeitamos** H_0 se $p \leq .05$
  - No exemplo, $p = .20$, então **não** rejeitamos H_0, ou seja, não podemos descartar que o parâmetro μ seja igual a μ_0. Nossa estimativa não foi **estatisticamente significativa**
- Rejeitar H_0 **não significa** que "provamos" que H_a é verdadeira.

Exercício

Bussab e Morettin, p. 339 (adaptado)

Uma máquina automática enche pacotes de café segundo uma distribuição normal com média igual a 500g. Periodicamente, coletamos uma amostra de 16 pacotes para verificar se a produção está sob controle. Uma dessas amostras registrou média $\bar{y} = 492g$ e variância igual a 400.

Você pararia ou não a produção para regular a máquina?

Exercício

Bussab e Morettin, p. 339 (adaptado)

Uma máquina automática enche pacotes de café segundo uma distribuição normal com média igual a 500g. Periodicamente, coletamos uma amostra de 16 pacotes para verificar se a produção está sob controle. Uma dessas amostras registrou média $\bar{y} = 492g$ e variância igual a 400.

Você pararia ou não a produção para regular a máquina?

- $H_0 : \mu = 500g$ e $H_1 : \mu \neq 500g$.
- Temos $n = 16$ e $se = \sqrt{400/16} = 5$. Se H_0 for verdadeira, $\bar{x} \sim N(500, 25)$.
- A estatística t é $t = (492 - 500)/5 = -1.6$, de modo que o p-valor é $Pr(|t| \geq 1.6)$, com $df = 16 - 1 = 15$. No R:

```
2 * (1 - pt(1.6, df = 15))
```

```
## [1] 0.130445
```

Região crítica

Recomenda-se escolher um **nível de significância** α *ex-ante*.

Com isso, podemos tomar decisões examinando simplesmente o **t-score**.

Por exemplo: para α = .05, com *df* = 15, temos os valores críticos:

```
qt(0.025, df = 15)
qt(0.975, df = 15)
```

```
## [1] -2.13145
## [1] 2.13145
```

Rejeitaríamos H_0 se $|t| \geq 2.13145$ ou, equivalentemente, $\bar{x} \leq 489.34$ ou $\bar{x} \geq 510.66$.

Como $t = -1.6$, não podemos rejeitar H_0 a 95% de confiança (α = .05).

Testes de significância e ICs

Conclusões baseadas em testes de significância são **consistentes** com conclusões baseadas em intervalos de confiança.

- Se um teste diz que um determinado valor é "plausível" para o parâmetro, o IC dirá a mesma coisa.

Na prática:

- Se **rejeitamos** a hipótese nula H_0 em um teste bilateral em torno da média μ, isso significa que $p \leq \alpha$. Nesse caso, o IC para μ com nível de significância de $(1 - \alpha)$ **não contém** o valor de H_0 para μ.
- Se **não rejeitamos** a hipótese nula H_0 em um teste bilateral, então $p \geq \alpha$. Nesse caso, o IC para μ com nível de significância de $(1 - \alpha)$ **contém** o valor de H_0.

Testes unilaterais ou de uma cauda só

Até agora, vimos testes para detectar desvios em qualquer direção em relação a H_0. Mas a **hipótese alternativa** pode ser direcional:

$$H_a : \mu > \mu_0 \qquad\qquad H_a : \mu < \mu_0$$

O cálculo do **t-score** é o mesmo de antes, isto é, $t = (\bar{x} - \mu_0)/se$. O que muda é o cálculo do p-valor, que se torna, respectivamente:

$$Pr(t \geq t_\alpha) \qquad\qquad Pr(t < t_\alpha)$$

Robustez

O uso da estatística t assume que a distribuição da variável na população é normal, de modo que a distribuição amostral de $\bar{y}$ é normal mesmo em amostras pequenas.

Pelo TCL, conforme n aumenta esse pressuposto se torna menos importante. Ou seja, o método é **robusto** mesmo quando o pressuposto de normalidade é violado...

... com uma ressalva: testes unilaterais não funcionam muito bem se a variável na população for fortemente assimétrica (mas testes bilaterais funcionam)

(No caso, isso vale só para médias e algumas estatísticas específicas)

Exercício

Bussab e Morettin, p. 356

Um fabricante afirma que seus cigarros contêm não mais que 30mg de nicotina. Uma amostra de 25 cigarros fornece média de 31.5mg e desvio padrão de 3mg. No nível de 5%, os dados refutam ou não o fabricante?

Exercício

Bussab e Morettin, p. 356

Um fabricante afirma que seus cigarros contêm não mais que 30mg de nicotina. Uma amostra de 25 cigarros fornece média de 31.5mg e desvio padrão de 3mg. No nível de 5%, os dados refutam ou não o fabricante?

- As hipóteses são $H_0 : \mu = 30$ e $H_1 : \mu > 30$.
- Supondo normalidade, o valor crítico t_c é obtido quando $Pr(t > t_c) = 0.05$, com $df = 24$:

```
qt(.95, 24)
```

```
## [1] 1.710882
```

- O valor observado é $t = (31.5 - 30)/(3/\sqrt{25}) = 2.5$, com $df = 24$. O p-valor é $\hat{\alpha} = P(t > t_0) = 0.01$.
- Como $\hat{\alpha} < \alpha = .05$ (ou seja, $t_o > t_c$), **rejeitamos** H_0.

Exercício

Bussab e Morettin, p. 356

Podemos obter a mesma resposta simplesmente analisando os intervalos confiança. No caso:

$$IC = 31.5 \pm t \cdot se = 31.5 \pm 2.06(3/\sqrt{25}) = 31.5 \pm 1.236 \rightarrow (30.26, 32.74)$$

O t-score foi obtido por:

```
qt(.975, 24)
```

```
## [1] 2.063899
```

Observe que o IC é construído bilateralmente!

Etapas

1. Pressupostos
2. Hipóteses
3. Estatística de teste
4. Cálculo do p-valor
5. Conclusão

...mesma estrutura, só pequenas mudanças, semelhantes às diferenças que vimos na aula de ICs.

Aproximação normal

Para amotras grandes (com pelo menos 15-20 casos em cada categoria) e π perto de 0.50, a aproximação normal funciona tão bem para testes de hipótese quanto para construir ICs.

Nesses casos, a estatística de teste é o **z-score**, calculado por:

$$z = \frac{\hat{\pi} - \pi_0}{se_0} \qquad \text{com} \qquad se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$$

(Repare que, embora o teste seja com outra distribuição, a sua forma é a mesma de antes)

Exercício

Em uma pesquisa, 624 de 1200 respondentes disseram que o governo deveria aumentar impostos para melhorar os serviços públicos, enquanto os outros 576 disseram que o governo deveria reduzir os gastos com serviços. Supondo α = 0.05, a maioria é a favor ou contra o aumento de impostos?

Exercício

Em uma pesquisa, 624 de 1200 respondentes disseram que o governo deveria aumentar impostos para melhorar os serviços públicos, enquanto os outros 576 disseram que o governo deveria reduzir os gastos com serviços. Supondo α = 0.05, a maioria é a favor ou contra o aumento de impostos?

Temos $H_0 : \pi = .5$ e $H_a : \pi \neq .5$. No R:

```
library(DescTools)
se0 <- sqrt(.5 * (1 - .5) / 1200)
z <- (.52 - .5) / se0
pvalor <- 2 * (1 - pnorm(z))
print(z)
print(pvalor)
```

```
## [1] 1.385641
## [1] 0.1658567
```

Resultado alternativo

Assim como para ICs, a aproximação normal às vezes não funciona tão bem. Por isso, o `prop.test`do R usa como padrão o score de Wilson com correção de continuidade:

```
prop.test(624, 1200, p = 0.50)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  624 out of 1200, null probability 0.5
## X-squared = 1.8408, df = 1, p-value = 0.1749
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4912980 0.5485725
## sample estimates:
##    p
## 0.52
```

Efeito do tamanho da amostra

Aumentar o tamanho da amostra torna os IC mais estreitos. Qual o efeito sobre o teste de significância? Supondo constantes as proporções de 52% e 48%, para qual tamanho de amostra nós passaríamos a **rejeitar** a hipótese nula $H_0 : \mu = 0.50$. O que isso significaria na prática?

Efeito do tamanho da amostra

Aumentar o tamanho da amostra torna os IC mais estreitos. Qual o efeito sobre o teste de significância? Supondo constantes as proporções de 52% e 48%, para qual tamanho de amostra nós passaríamos a **rejeitar** a hipótese nula $H_0 : \mu = 0.50$. O que isso significaria na prática?

Pela aproximação normal, rejeitamos H_0 se

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{.5(1-.5)}{n}}} \geq z_{critico} = 1.96$$

Logo:

$$z_{critico} = \frac{.52 - .5}{.5/\sqrt{n}} = (.02/.5)\sqrt{n} \qquad \rightarrow \qquad n = \left(\frac{z_{critico}}{0.04}\right)^2$$

Efeito do tamanho da amostra

```
# resolvendo
z_critico <- qnorm(.975)
n <- (z_critico / .04)^2
print(n)
# Conferindo com aproximacao normal
se0 <- sqrt( (.5 * (1 - .5)) / n)
z <- (.52 - .50) / se0
print(pnorm(-z) + (1 - pnorm(z)))
# Da certo com score de Wilson?
prop.test( .52*n, n, p = .5)$p.value
```

```
## [1] 2400.912
## [1] 0.05
## [1] 0.05243374
```

Efeito do tamanho da amostra

```
# Podemos aumentar aos poucos, por tentativa e erro
n <- n + 40
prop.test( .52*n, n, p = .5)$p.value
n <- n + 9
prop.test( .52*n, n, p = .5)$p.value
n <- n + 1
prop.test( .52*n, n, p = .5)$p.value
# Agora sim!
print(n)
```

```
## [1] 0.05046719
## [1] 0.05003555
## [1] 0.04998783
## [1] 2450.912
```

Erro tipo 1 e erro tipo 2

Rejeitamos H_0 se $p \leq \alpha$ para um α pré-especificado; caso contrário, não rejeitamos.

- A boa prática é sempre reportar o p-valor; não basta dizer se é estatisticamente significativo ou não.

Mas o que isso significa?

- Devido à variabilidade amostral, nenhuma decisão é determinística. Há sempre incerteza. O α que escolhemos reflete o grau de incerteza e erros em potencial que estamos dispostos a tolerar.

Erro tipo 1 e erro tipo 2

Erro tipo 1

Ocorre quando H_0 é **verdadeira**, mas nós erroneamente **rejeitamos** ela.

O valor do nível de significância α determina a probabilidade de erro tipo 1.

- se $\alpha = 0.05$, então vamos rejeitar erroneamente H_0 em 5% dos casos em amostras repetidas com o mesmo tamanho.

Erro tipo 2

Ocorre quando H_0 é **falsa**, mas erroneamente **não é rejeitada**.

- A probabilidade de erro tipo 2 é dada por β, cuja determinação é mais difícil porque normalmente não especificamos valores fixos para o parâmetro sob a hipótese alternativa.

Trade-off entre tipos de erros

Reduzir o risco de um tipo de erro aumenta o risco do outro

Suponha um julgamento criminal em que H_0 é que o réu é inocente e H_1 é que ele culpado. Os jurados rejeitam H_0 e condenam o réu se a evidência for forte o suficiente.

- Erro tipo 1: condenação de um réu inocente
- Erro tipo 2: não condenação de um réu culpado

O que acontece se os jurados quiserem minimizar a chance de um erro tipo 1? Na prática, vão exigir mais evidências para rejeitar H_0. Somente os casos com evidências extraordinariamente fortes serão condenados...

...com isso, quase nenhum inocente irá para a cadeia, mas, na prática, o júri vai acabar absolvendo também muitos culpados.

Trade-off entre tipos de erros

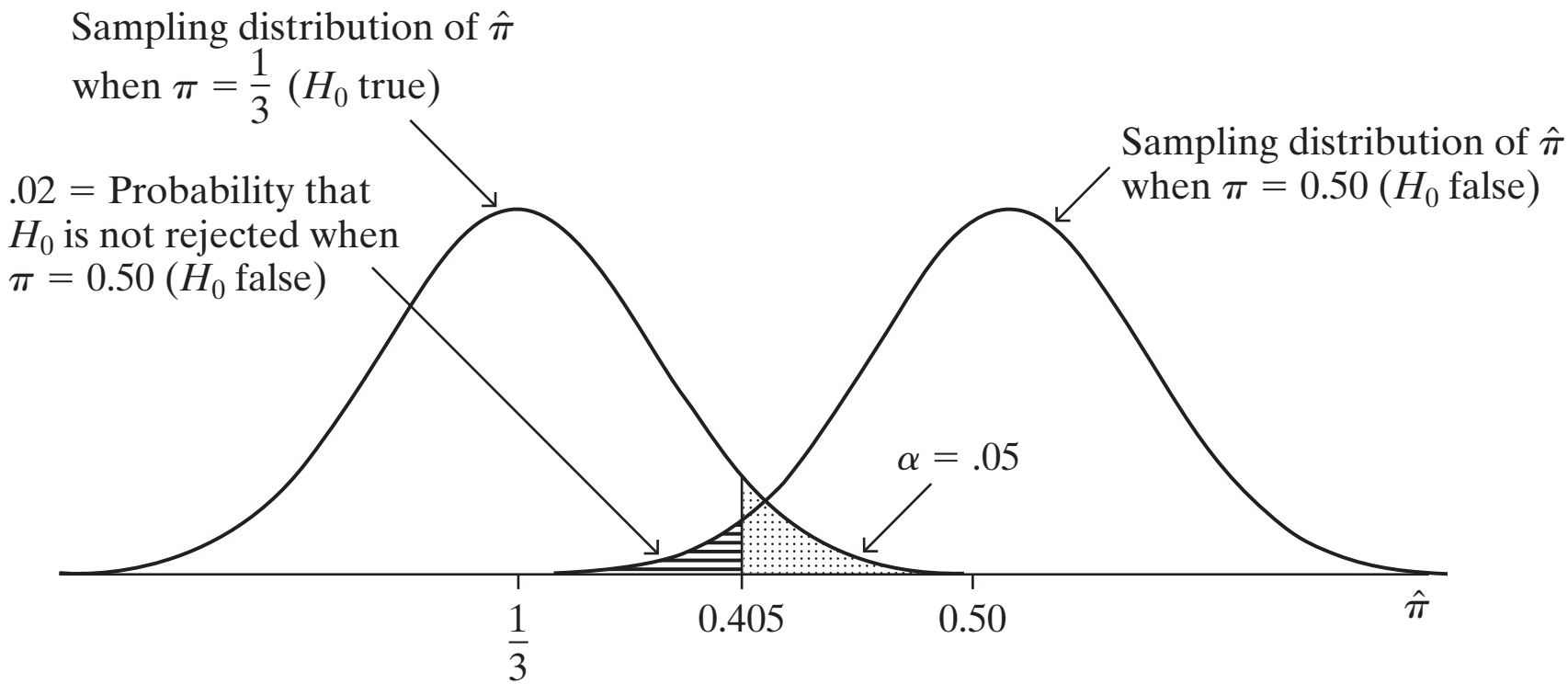
Testes de hipótese funcionam do mesmo jeito: se adotarmos valores muito pequenos para α, raramente vamos rejeitar de modo errôneo a hipótese nula, mas vamos errar com mais frequência ao não rejeitarmos hipóteses nulas falsas.

O único jeito de reduzir simultaneamente α e β é aumentando o **tamanho da amostra**.

Na prática, tipicamente escolhemos só o valor de α, que está sob nosso controle.

- Se H_0 for verdadeira, a probabilidade de um erro do tipo 1 é α.

Exemplo


Sampling distribution of $\hat{\pi}$ when $\pi = \frac{1}{3}$ (H_0 true)
.02 = Probability that H_0 is not rejected when $\pi = 0.50$ (H_0 false)
Sampling distribution of $\hat{\pi}$ when $\pi = 0.50$ (H_0 false)
$\alpha = .05$
$\frac{1}{3}$
0.405
0.50
$\hat{\pi}$

Poder de um teste

A probabilidade de rejeitar **corretamente** H_0 é chamada de **poder** de um teste.

Para um valor específico do parâmetro no intervalo delimitado por H_a, o poder é:

$$\text{Poder} = 1 - Pr(\text{Erro Tipo 2}) = 1 - \beta$$

O poder diz respeito à capacidade de um teste de captar diferenças de determinada magnitude entre um valor de interesse e a hipótese nula.

- Idealmente, testes de significância estatística também deveriam sempre reportar o poder, especialmente quando n é pequeno.

Tipos de amostras

Mulheres gastam mais tempo do que homens em tarefas domésticas?

Esse tipo de pergunta de pesquisa implica a comparação entre médias de dois grupos, o que exige tanto o cálculo de estimativas de ponto para cada grupo quanto a realização de testes de significância da diferença entre as médias.

Esses testes são desdobramentos do que vimos até aqui, mas com um detalhe importante: os **erros padrão** para as estatísticas de teste dependem do **tipo de amostra** que usamos.

- Independentes ou dependentes (pareadas)

Calculando os erros padrão

Para comparar médias em duas populações, consideramos que o parâmetro de interesse é $\mu_2 - \mu_1$ e podemos estimá-lo pelas médias amostrais $\bar{y_2} - \bar{y_1}$.

A **distribuição amostral** de $\bar{y_2} - \bar{y_1}$ tem valor esperado igual a $\mu_2 - \mu_1$ e, para **amostras grandes**, tem distribuição (aproximadamente) normal, conforme o TCL.

O erro padrão da distribuição amostral de $\bar{y_2} - \bar{y_1}$ descreve quão precisamente $\bar{y_2} - \bar{y_1}$ funciona como estimador de $\mu_2 - \mu_1$.

- O erro padrão descreve a variabilidade de estimativa de diferentes estudos potenciais com amostras de mesmo tamanho.

Calculando os erros padrão

Amostras independentes

Para duas amostras independentes, o erro padrão de $\bar{y_2} - \bar{y_1}$ é dado por:

$$se = \sqrt{se_1^2 + se_2^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

... ou seja, o erro padrão da diferença é maior do que os erros padrão individuais.

Amostras dependentes

Nesse caso, construímos $y_{dif} = y_{2i} - y_{1i}$ e calculamos $\bar{y_{dif}}$, pois a média das diferenças é igual à diferença entre médias. Para o erro padrão, a fórmula é a mesma que vimos para uma média simples:

Intervalos de confiança

Para diferenças ou razões para médias e proporções, a construção dos intervalos de confiança segue o que vimos anteriormente:

- A estimativa de ponto é a diferença ou razão
- O erro padrão depende se as amostras forem independentes ou não
- Para médias de variáveis contínuas, os graus de liberdade para os t-score são complicadinhos, mas o R faz tudo

Quando o IC de uma **diferença** entre médias (ou proporções) contém **zero**, não podemos descartar a hipótese nula de que as duas médias (proporções) são idênticas.

Testes de significância

Proporções

Para amostras independentes, o erro padrão sob a hipótese nula $H_0 : \pi_1 = \pi_2$ é dado por $se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1} + \frac{1}{n_2})}$, em que $\hat{\pi}$ é a proporção combinada das duas amostras.

Médias

O método básico é semelhante.

Vejam Agresti 2018, cap. 7 para mais informações

Exemplo

Mulheres gastam mais tempo do que homens em tarefas domésticas?

Dados de Agresti 2018, p. 179, do GSS 2012 $\rightarrow$ Horas gastas por semana

	N	Média	SD
Homens	583	8.3	9.4
Mulheres	693	11.9	12.7

Exemplo

Mulheres gastam mais tempo do que homens em tarefas domésticas?

Dados de Agresti 2018, p. 179, do GSS 2012 $\rightarrow$ Horas gastas por semana

	N	Média	SD
Homens	583	8.3	9.4
Mulheres	693	11.9	12.7

- Temos $H_0 : \mu_{homens} = \mu_{mulheres}$ e $H_1 : \mu_{homens} \neq \mu_{mulheres}$
- Estimativa de ponto: $11.9 - .8.3 = 3.6$
- Erro padrão: $se = \sqrt{\frac{9.4^2}{583} + \frac{12.7^2}{693}} = 0.62$

Exemplo

Mulheres gastam mais tempo do que homens em tarefas domésticas?

Intervalo de confiança a 95%

Como o tamanho das amostras é grande, podemos usar o **z-score** no lugar do **t-score**:

$$3.6 \pm 1.96se = 3.6 \pm 1.96 \cdot 0.62 = 3.6 \pm 1.2 \qquad \rightarrow \qquad (2.4, 4.8)$$

Teste de significância

$$t = \frac{(\bar{y_2} - \bar{y_1}) - 0)}{se} = \frac{3.6 - 0}{0.62} = 5.8 \qquad \rightarrow \qquad p \leq 0.0001$$

Exemplo

```
library(tidyverse); library(nycflights13)
jfk <- flights %>% filter(origin == "JFK")
ewr <- flights %>% filter(origin == "EWR")
t.test(jfk$arr_delay, ewr$arr_delay, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  jfk$arr_delay and ewr$arr_delay
## t = -18.826, df = 225780, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal
## 95 percent confidence interval:
##  -3.925750 -3.185397
## sample estimates:
## mean of x mean of y
##  5.551481  9.107055
```

Exercício

Aplique um teste t **dependente** comparando `dep_delay` e `arr_delay` em `flights`. O que significa? (Dica: use a opção `paired = TRUE`)

Exercício

Aplique um teste t **dependente** comparando `dep_delay` e `arr_delay` em `flights`. O que significa? (Dica: use a opção `paired = TRUE`)

```
t.test(flights$dep_delay, flights$arr_delay, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  flights$dep_delay and flights$arr_delay
## t = 179.46, df = 327345, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to
## 95 percent confidence interval:
##  5.597967 5.721591
## sample estimates:
## mean difference
##        5.659779
```

Exercício

Agresti 2018, p. 183: oração ajuda os pacientes de cirurgias coronarianas?

Grupo	Com complicações	Sem complicações	Total
Tratamento	315	289	604
Controle	304	293	597

Exercício

Agresti 2018, p. 183: oração ajuda os pacientes de cirurgias coronarianas?

Grupo	Com complicações	Sem complicações	Total
Tratamento	315	289	604
Controle	304	293	597

```
# IC com aprox normal
pi1 <- 315/604
pi2 <- 304/597
se1 <- sqrt( pi1 * (1 - pi1) / 604)
se2 <- sqrt( pi2 * (1 - pi2) / 597)
se <- sqrt( se2^2 + se1^2)
print(c(pi2 - pi1 + qnorm(.025)*se, pi2 - pi1 + qnorm(.975)*se))
```

```
## [1] -0.06883625  0.04421536
```

Exercício

```
# IC e teste alternativo
prop.test(x = c(304, 315), n = c(597, 604), correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##
## data:  c(304, 315) out of c(597, 604)
## X-squared = 0.18217, df = 1, p-value = 0.6695
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.06883625  0.04421536
## sample estimates:
##    prop 1    prop 2
## 0.5092127 0.5215232
```

Três limitações

1. Significância estatística não implica significância prática

Um p-valor pequeno (e.g., $p = 0.001$) é altamente significativo em termos estatísticos, mas não necessariamente indica que se trata de um achado importante em termos substantivos.

2. ICs costumam ser mais úteis que testes de significância

Um teste apenas indica se um valor particular de H_0 é plausível, mas não nos diz quais outros valores em potencial são plausíveis. ICs fazem isso.

3. P-valores podem nos induzir a erros

O p-valor não nos diz qual a probabilidade de que H_0 seja verdadeira e significância estatística às vezes ocorre por acaso. Nunca maltrate os dados até obter algo significativo!

Recapitulação

Introdução

Testes de hipóteses para médias

Mais propriedades importantes

Testes de significância para proporções

Tipos de erros em testes

Comparações entre médias

Limitações de testes de significância

Próxima aula

Próxima aula

Atividade

O prazo para entrega da atividade #5 é 21/11

Leituras obrigatórias

Agresti 2018, cap.9 a 10 6

Leituras optativas

Bussab e Morettin 2010 cap. 16

Huntington-Klein, cap. 13