Métodos Quantitativos

Aula 07. Inferência estatística

Pedro H. G. Ferreira de Souza

pedro.ferreira@ipea.gov.br

Mestrado Profissional em Políticas Públicas e Desenvolvimento

Instituto de Pesquisa Econômica Aplicada (Ipea)

07 nov. 2022

Aula passada

Amostragem

Viés amostral (ou de seleção), aleatorização, sorteio de AAS

Fundamentos de probabilidade

Espaço amostral, regras básicas, probabilidade conjunta, probabilidade condicional, independência

Variáveis aleatórias

Discretas e contínuas, distribuições uniforme, Bernoulli e normal

Distribuições amostrais

Estatística amostral como variável aleatória que possui uma distribuição de probabilidade, erro padrão como desvio padrão da distribuição amostral

Aula passada

Teorema Central do Limite

Independentemente da distribuição de x, a distribuição amostral da média amostral $\bar{x}$ é (aproximadamente) normal com parâmetros:

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Ou seja, a variabilidade da média depende do desvio padrão de x na população e do tamanho n da amostra.

Distribuição amostral de outras estatísticas

O TCL pode ser estendido para outras estatísticas, mas não é válido para todas.

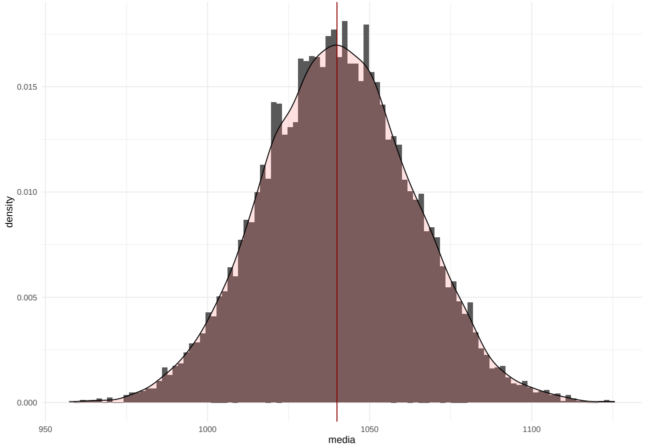
Aula passada

Simulação da distribuição amostral da média com 10,000 amostras e $n = 1,000$

```
library(tidyverse)
library(nycflights13)
dist <- as.vector(flights$distance)
amostras <- replicate(10000, mean(sample(dist, 1000)))
amostras <- data.frame(media = amostras)
ggplot(amostras, aes(x = media)) +
  geom_histogram(aes(y=..density..), bins = 100) +
  geom_density(alpha = .2, fill = 'indianred1') +
  geom_vline(aes(xintercept = mean(flights$distance)),
             color='darkred') +
  theme_minimal()
```

Aula passada

Simulação da distribuição amostral da média com 10,000 amostras e $n = 1,000$



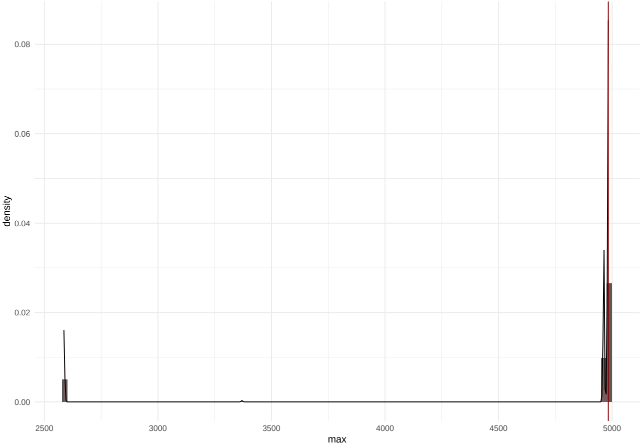
Bônus

Simulação da distribuição amostral do máximo com 10,000 amostras e $n = 1,000$

```
library(tidyverse)
library(nycflights13)
dist <- as.vector(flights$distance)
amostras <- replicate(10000, max(sample(dist, 1000)))
amostras <- data.frame(max = amostras)
ggplot(amostras, aes(x = max)) +
  geom_histogram(aes(y=..density..), bins = 100) +
  geom_density(alpha = .2, fill = 'indianred1') +
  geom_vline(aes(xintercept = max(flights$distance)),
             color='darkred') +
  theme_minimal()
```

Bônus

Simulação da distribuição amostral do máximo com 10,000 amostras e $n = 1,000$



O que é inferência estatística?

Definição

Inferência estatística é o processo de usar **dados amostrais** para estimar **parâmetros populacionais**, isto é, fazer generalizações sobre uma população a partir de uma amostra.

Como nossa amostra (aleatória) é somente uma de muitas amostras possíveis, ou seja, como há **flutuação amostral**, nossas estimativas se desdobram em dois componentes:

- Uma **estimativa de ponto** é o número que representa nosso melhor palpite para o parâmetro de interesse
- Uma **estimativa de intervalo** ou **intervalo de confiança** em torno da estimativa de ponto quantifica nossa incerteza quanto ao valor exato do parâmetro

Estimadores e estimativas

Um **estimador** é a fórmula ou "receita" aplicada aos dados para produzir **estimativas**, isto é, para gerar palpites sobre os parâmetros populacionais desconhecidos.

É impossível justificar uma estimativa por si só, afinal, não sabemos o número real. A justificativa é sempre sobre o estimador.

- A "aceitabilidade" de uma estimativa computada em uma amostra depende da "aceitabilidade" do método de estimação (estimador)
- Um estimador T de um parâmetro θ é qualquer função das observações da amostra, ou seja, $T = g(x_1, x_2, \ldots, x_n)$
- O problema central, portanto, é escolher uma função $g(.)$ que gere estimativas "próximas" de θ segundo algum critério

Exemplos de estimadores (i)

Até aqui, estimamos parâmetros populacionais "imitando" na amostra o que acontece na população: por ex., usamos $\bar{x}$ para estimar μ. Mas por que isso é válido?

Estimadores de momentos

A média populacional é o **primeiro momento** da distribuição, ou seja, $\mu_1 = E(X)$. Generalizando, o k-ésimo momento é dado por $\mu_k = E(X^k)$.

A estimação pelo **métodos dos momentos** é feita quando igualamos os k primeiros momentos teóricos aos respectivos momentos amostrais e resolvemos.

Grosso modo, estimadores de moemntos são consistentes, mas às vezes enviesados.

Exemplos de estimadores (ii)

Estimadores de máxima verossimilhança (MLE)

São estimadores que maximizam a probabilidade de obtermos a amostra particular observada, ou seja, estimam os parâmetros populacionais que tornam nossa amostra a "mais provável".

Matematicamente, no caso da média populacional, o MLE também é a média amostral.

R. A. Fisher desenvolveu essa classe de estimadores, mostrando que, para amostras grandes, eles são eficientes, consistentes e têm distribuição amostral aproximadamente normal.

Intervalo de confiança

A estimativa de intervalo ou intervalo de confiança, por sua vez, depende tanto da nossa **estimativa de ponto** quanto da **distribuição amostral** dessa estimativa de ponto.

Frequentemente, a distribuição amostral é aproximadamente normal. Como vimos na última aula, é bastante simples quantificar a incerteza nesse tipo de distribuição. Afinal, em uma distribuição normal padrão $Z \sim N(0, 1)$:

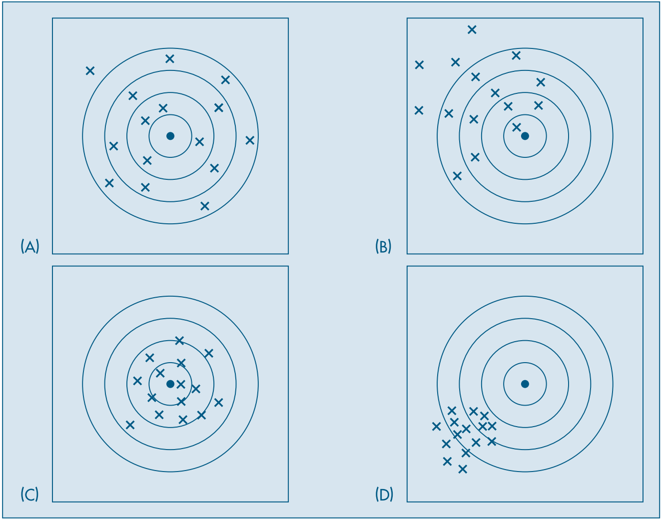
- $\Pr(-1 \leq z \leq 1) \approx 68\%$
- $\Pr(-1.96 \leq z \leq 1.96) \approx 95\%$
- $\Pr(-3 \leq z \leq 3) \approx 99.7\%$

Pacotes

Instalem (se necessário) e carreguem os pacotes que vamos usar hoje:

```
library(boot)
library(tidyverse)
library(summarytools)
library(DescTools)
library(nycflights13)
```

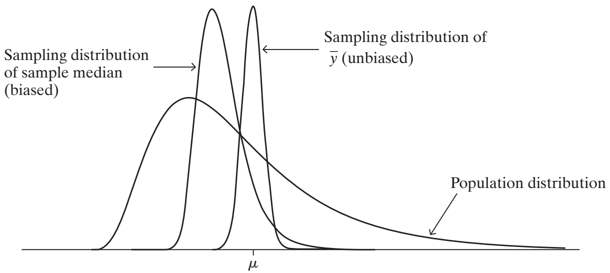
Como avaliar um estimador?



Ausência de viés

Um estimador é **não viesado** se a média de sua distribuição amostral for igual ao parâmetro de interesse.

Ou seja, se repetirmos o "experimento" infinitas vezes, calcularmos o valor do estimador a cada vez e, no fim, fizermos a média de nossas estimativas, essa média será igual a θ.



Ausência de viés

Formalmente, um estimador é não viesado se, para todo θ:

$$E(T) = \theta$$

Portanto, o **viés** de T é dado por $V(T) = E(T) - \theta$.

Vimos anteriormente que a **média amostral** $\bar{x}$ é um estimador não viesado de μ e que a **proporção amostral** $\hat{p}$ é um estimador não viesado de p.

Bussab e Morettin (2010: p. 299-300) explicam por que é preciso o denominador $n - 1$ para obter um estimador não viesado para a variância:

$$E(s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2) = \sigma^2$$

Consistência

Um estimador é consistente se as estimativas "convergem" para o valor real do parâmetro θ conforme o tamanho da amostra aumenta, isto é, sua distribuição amostral torna-se crescentemente concentrada em torno de θ:

$$lim_{n \to \infty} E(T_n) = \theta$$

$$lim_{n \to \infty} Var(T_n) = 0$$

Ausência de viés e consistência não necessaramente andam juntas:

- A média amostral $\bar{x}$ é um estimador não viesado e consistente de μ
- A variância amostral sem correções é um estimador viesado, porém consistente de σ^2

Eficiência

Dada a ausência de viés, outra propriedade desejável para um estimador é que "na média" ele produza estimativas mais próximas ao parâmetro populacional do que opções alternativas.

- Um estimador **eficiente** é um estimador não-viesado que tem **erro padrão** menor do que o de todos os outros estimadores não-viesados.
- Formalmente, se T e T' são estimadores não viesados de um mesmo parâmetro θ, T é mais eficiente do que T' se $Var(T) < Var(T')$

Eficiência

Exemplo

Considere uma variável $X \sim N(\mu, \sigma^2)$, ou seja, a média e a mediana populacionais são iguais. Sejam $\bar{x}$ e md a média e a mediana em uma amostra de tamanho n, qual dos dois é o melhor para estimar a mediana populacional?

- Bussab e Morettin (2010, p. 302) mostram que os dois são estimadores não viesados, mas $\bar{x}$ é mais eficiente, pois:

$$\frac{Var(md)}{Var(\bar{x})} = \frac{\pi}{2} > 1$$

Estimando médias e desvios padrão

É comum, mas não necessário, usarmos estatísticas análogas amostrais para estimar um parâmetro populacional

- A média e a proporção na amostra são estimadores não viesados e eficientes de suas contrapartes populacionais
- A variância amostral $\frac{1}{n}\sum(x_i - \bar{x})^2$ é um estimador enviesado, porém consistente da variância populacional.
- A maioria dos *softwares* automaticamente calcula $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$, que é um estimador não viesado, eficiente e consistente.

Bônus: erro quadrado médio (MSE)

O erro amostral que cometemos ao estimar θ por T baseado em uma amostra é dado por $e = T - \theta$. Assim, o **erro quadrado médio** é:

$$MSE(T;\theta) = E(e^2) = E(T - \theta)^2 = Var(T) + V^2$$

Para um estimador não viesado, o MSE é simplesmente a variância do estimador.

O MSE é uma medida de qualidade do estimador muito usada em modelos mais complexos. Afinal, em alguns casos preferimos podemos preferir um estimador viesado, porém com baixa variância a um estimador não viesado com variância enorme.

O que são ICs?

Nossa amostra é apenas uma de muitas possíveis e, portanto, nossas estimativas de ponto nunca serão (na prática) 100% precisas.

ICs quantificam essa incerteza, apontando uma **margem de erro** calculada a partir de um **grau de confiança** escolhido:

$$\text{IC} = \text{estimativa de ponto} \pm \text{margem de erro}$$

O grau de confiança é a probabilidade de que esse método produza um intervalo que efetivamente contenha o parâmetro.

O tamanho da margem de erro depende da distribuição amostral do estimador de ponto.

Mais sobre a margem de erro

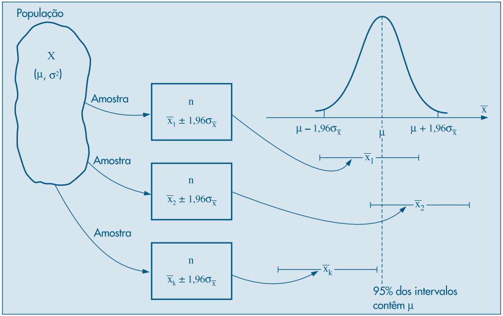
A margem erro é tipicamente dada por $z \cdot se$ ou $t \cdot se$, ou seja, ela resulta da multiplicação do z-score ou t-score associado ao **nível de confiança** escolhido (por hábito, 95%) por uma estimativa do **erro padrão** *se* da distribuição amostral do estimador.

- Quanto **maior** o grau de confiança, **maior** a margem de erro
  - Intuitivamente, *ceteris paribus*, ICs a 99% são mais "largos" que ICs a 95%, que são mais "largos" do que ICs a 90%, e assim por diante.
- Quanto **maior** o tamanho da amostra, **menor** a margem de erro
  - Vimos isso na aula passada no exemplo sobre pesquisas eleitorais e retornaremos a esse caso mais adiante.

Como interpretar um IC

Suponha que estimamos a média μ com nível de confiança de 95%:

Figura 11.3: Significado de um IC para μ, com $\gamma = 0{,}95$ e σ^2 conhecido.



μ não é uma variável aleatória, mas um parâmetro fixo. O que o IC nos diz nesse caso é que, **em zilhões de amostras independentes repetidas**, em 95% das estimativas o IC estimado vai conter o parâmetro μ

Propriedades importantes

1. O IC aumenta conforme o grau de confiança aumenta e diminuiu conforme o tamanho da amostra aumenta.
2. A probabilidade de erro α é a probabilidade de que o IC **não** contenha o parâmetro, dada por **1 - nível de confiança**, tipicamente com $\alpha = 1 - 0.95 = 5\%$ (mas é só um valor convencional, não há nada de especial nele)
3. O valor de z para o IC corresponde, para um teste de duas caudas, ao z associado a $\alpha/2$ e $1 - \alpha/2$ na distribuição normal padrão.
4. O IC descreve o desempenho do método em incontáveis amostras repetidas; um IC específico pode conter ou não o parâmetro.
5. O IC depende **criticamente** de quão bem a distribuição amostral se aproxima de uma distribuição normal.

IC para proporções: aproximação normal (i)

Seja π uma proporção populacional, o que, por definição, implica $0 \leq \pi \leq 1$, e seja $\hat{\pi}$ a proporção amostral.

Uma proporção pode ser modelada como uma distribuição binomial $B(n, p)$. Mas se codificarmos "sucessos" como 1 e "fracassos" como 0, a proporção equivale à média da VA e, como vimos, a distribuição amostral da média é **aproximadamente** normal com média e erro padrão:

$$\mu = \pi$$

$$\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

IC para proporções: aproximação normal (ii)

Aprendemos que em uma distribuição normal 95% da área está entre ± 1.96 desvios padrão da média...

... logo, podemos construir o **IC com nível de confiança de 95%** para a proporção $\hat{\pi}$:

$$\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$$

IC para proporções: aproximação normal (ii)

Aprendemos que em uma distribuição normal 95% da área está entre ± 1.96 desvios padrão da média...

... logo, podemos construir o **IC com nível de confiança de 95%** para a proporção $\hat{\pi}$:

$$\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$$

Oh-oh, mas não conhecemos $\sigma_{\hat{\pi}}$ porque não conhecemos o parâmetro populacional π (só $\hat{\pi}$, nossa estimativa amostral).

E agora?

IC para proporções: aproximação normal (iii)

Para construir o IC, precisamos estimar também o erro padrão, usando a proporção amostral no lugar do parâmetro populacional:

$$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Logo, o **intervalo de confiança a 95%** para π é:

$$\hat{\pi} \pm 1.96 se = \hat{\pi} \pm 1.96 \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Exercício

Suponha que o DataPedro entrevistou 7740 eleitores, sendo que 4024 declararam voto no candidato A e 3716 disseram que vão votar em B. Calcule o IC para o candidato A a 90%, 95% e 99%.

Exercício

Suponha que o DataPedro entrevistou 7740 eleitores, sendo que 4024 declararam voto no candidato A e 3716 disseram que vão votar em B. Calcule o IC para o candidato A a 90%, 95% e 99%.

```
# Dados
N <- 7740
pA <- 4024 / N
se = sqrt( (pA*(1 - pA)) / N )
# IC 90%
(z90 <- qnorm(.95))
(ic90 <- c(pA - z90*se, pA, pA + z90*se) * 100)
```

Exercício

Suponha que o DataPedro entrevistou 7740 eleitores, sendo que 4024 declararam voto no candidato A e 3716 disseram que vão votar em B. Calcule o IC para o candidato A a 90%, 95% e 99%.

```
# Dados
N <- 7740
pA <- 4024 / N
se = sqrt( (pA*(1 - pA)) / N )
# IC 90%
(z90 <- qnorm(.95))
(ic90 <- c(pA - z90*se, pA, pA + z90*se) * 100)
```

```
## [1] 1.644854
```

```
## [1] 51.05559 51.98966 52.92374
```

Exercício

Suponha que o DataPedro entrevistou 7740 eleitores, sendo que 4024 declararam voto no candidato A e 3716 disseram que vão votar em B. Calcule o IC para o candidato A a 90%, 95% e 99%.

Exercício

Suponha que o DataPedro entrevistou 7740 eleitores, sendo que 4024 declararam voto no candidato A e 3716 disseram que vão votar em B. Calcule o IC para o candidato A a 90%, 95% e 99%.

```
# Dados
N <- 7740
pA <- 4024 / N
se = sqrt( (pA*(1 - pA)) / N )
# IC 95% e 99%
(ic95 <- c(pA+qnorm(.025)*se, pA+qnorm(.975)*se) * 100)
(ic99 <- c(pA+qnorm(.005)*se, pA+qnorm(.995)*se) * 100)
```

Exercício

Suponha que o DataPedro entrevistou 7740 eleitores, sendo que 4024 declararam voto no candidato A e 3716 disseram que vão votar em B. Calcule o IC para o candidato A a 90%, 95% e 99%.

```
# Dados
N <- 7740
pA <- 4024 / N
se = sqrt( (pA*(1 - pA)) / N )
# IC 95% e 99%
(ic95 <- c(pA+qnorm(.025)*se, pA+qnorm(.975)*se) * 100)
(ic99 <- c(pA+qnorm(.005)*se, pA+qnorm(.995)*se) * 100)
```

```
## [1] 50.87664 53.10269
```

```
## [1] 50.52691 53.45242
```

Margem de erro

A margem de erro depende de. . . .

- O **grau de confiança** escolhido, que, por sua vez, determina z
- O **erro padrão** estimado $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$, que depende de $\hat{\pi}$ e de n

```
#  IC 95%
pA <- 4024 / 7740
se = sqrt( (pA*(1 - pA)) / 7740 )
(ic <- c(pA+qnorm(.025)*se, pA, pA+qnorm(.975)*se) * 100)
(margem_de_erro <- c(-qnorm(.025)*se, qnorm(.975)*se))
```

```
## [1] 50.87664 51.98966 53.10269
## [1] 0.01113021 0.01113021
```

Margem de erro

- Quanto **maior** o grau de confiança, **maior** a margem de erro
- Quanto **maior** o tamanho da amostra, **menor** a margem de error

Na prática, não controlamos $\hat{\pi}$, mas podemos escolher a maior margem de erro que estamos dispostos a tolerar se formos conservadores quanto a $\hat{\pi}$ e ajustarmos o tamanho da amostra ao grau de confiança desejado.

Supondo $\hat{\pi} = 0.50$, para uma margem de erro de até e, o tamanho da amostra n tem que ser:

$$e = z\sqrt{\frac{0.5(1-0.5)}{n}} \rightarrow n = \frac{0.25z^2}{e^2} = n = 0.25\left(\frac{z}{e}\right)^2$$

IC para proporções: aproximação normal (iv)

A aproximação normal funciona bem quando a amostra é “grande” e o parâmetro populacional está mais perto de 0.5 do que de zero ou 1. Em contrapartida, às vezes dá muito errado. Exemplo claro:

```
n <- 100
p <- 2 / n
(raro95pct <- c(p + qnorm(.025)*sqrt(p * (1-p) / n),
                p,
                p + qnorm(.975)*sqrt(p * (1-p) / n)) * 100)
(freq95pct <- c((1-p) + qnorm(.025)*sqrt(p * (1-p) / n),
                (1-p),
                (1-p) + qnorm(.975)*sqrt(p * (1-p) / n)) * 100)
```

```
## [1] -0.7439496  2.0000000  4.7439496
## [1]  95.25605  98.00000 100.74395
```

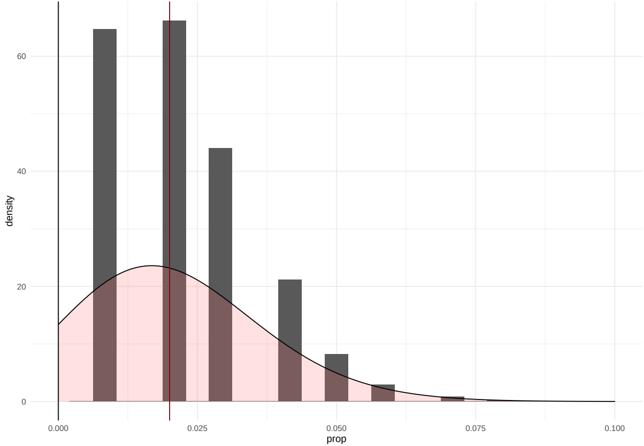
IC para proporções: aproximação normal (v)

O que deu errado? Podemos visualizar simulando a distribuição amostral. Vamos supor que $\mu = 2\%$:

```
sim_da <- replicate(20000,
                    mean(rbinom(n=100, size=1, prob=.02))) %>%
            data.frame(prop = .)
ggplot(sim_da, aes(x = prop)) +
  geom_histogram(aes(y=..density..), bins = 25) +
  geom_density(bw = .01, alpha = .2, fill = 'indianred1') +
  geom_vline(aes(xintercept = .02),
             color='darkred') +
  geom_vline(aes(xintercept = 0),
             color='black') +
  xlim(0, .1) +
  theme_minimal()
```

IC para proporções: aproximação normal (v)

O que deu errado? Podemos visualizar simulando a distribuição amostral. Vamos supor que μ = 2%:



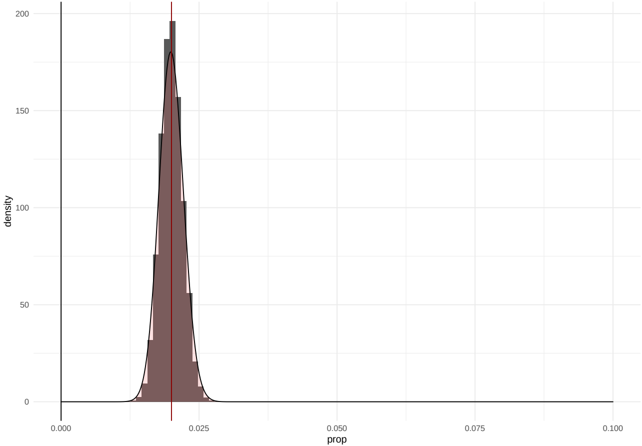
IC para proporções: aproximação normal (vi)

Se a amostra fosse muito maior, não haveria tanto problema. Vamos refazer agora com $n = 5000$, o que gera um IC $(1.61\%, 2.39\%)$:

```
sim_da <- replicate(20000,
                    mean(rbinom(n=5000, size=1, prob=.02))) %>%
            data.frame(prop = .)
ggplot(sim_da, aes(x = prop)) +
  geom_histogram(aes(y=..density..), bins = 100) +
  geom_density(bw = .001, alpha = .2, fill = 'indianred1') +
  geom_vline(aes(xintercept = .02),
             color='darkred') +
  geom_vline(aes(xintercept = 0),
             color='black') +
  xlim(0, .1) +
  theme_minimal()
```

IC para proporções: aproximação normal (vi)

Se a amostra fosse muito maior, não haveria tanto problema. Vamos refazer agora com $n = 5000$, o que gera um IC (1.61%, 2.39%):



Uma opção melhor para IC de proporções

O *score de Wilson com correção de continuidade* faz ajustes na fórmula do IC. No R é só executar `prop.test(x, n, conf.level == XX)`

Uma opção melhor para IC de proporções

O *score de Wilson com correção de continuidade* faz ajustes na fórmula do IC. No R é só executar `prop.test(x, n, conf.level == XX)`

```
# Exemplo eleicoes
prop.test(4024, 7740)$conf.int[1:2]
# Exemplo raro a 95% com n = 5000
prop.test(100, 5000)$conf.int[1:2]
# Exemplo raro a 95% e 99% com n = 100
prop.test(2, 100)$conf.int[1:2]
prop.test(2, 100, conf.level = .99)$conf.int[1:2]
```

```
## [1] 0.5086947 0.5310787
## [1] 0.01638153 0.02437436
## [1] 0.003471713 0.077363988
## [1] 0.002398955 0.103443650
```

Exercício

Uma empresa quer lançar um novo refrigerante. Para isso, fazem um experimento cego em que os entrevistados dão um gole e depois dizem se gostaram ou não.

Pelo método da aproximação normal, qual deve ser o n se quisermos um IC com amplitude de até **1pp** a **99%** de confiança? O que acontece quando usamos o mesmo n para calcular o IC via *score de Wilson*?

Exercício

Uma empresa quer lançar um novo refrigerante. Para isso, fazem um experimento cego em que os entrevistados dão um gole e depois dizem se gostaram ou não.

Pelo método da aproximação normal, qual deve ser o n se quisermos um IC com amplitude de até **1pp** a **99%** de confiança? O que acontece quando usamos o mesmo n para calcular o IC via *score de Wilson*?

```
# Aproximacao normal conservadora
(n <- 0.25 * (qnorm(.995)/.005)^2)
(c(.5 + qnorm(.005)*.5/sqrt(n), .5 + qnorm(.995)*.5/sqrt(n)))
# Score de Wilson
prop.test(n/2, n, conf.level = .99)$conf.int[1:2]
```

```
## [1] 66348.97
## [1] 0.495 0.505
## [1] 0.4950002 0.5049998
```

IC para proporções multinomiais (i)

Cálculo do IC é bem mais complicado – há muitos métodos disponíveis, desde a aproximação normal (não recomendado) até estimação simultânea. No R, usamos o comando `MultinomCI`, do pacote `DescTools`.

```
# Pesquisa com varios candidatos
pesq <- data.frame(cand = c('A', 'B', 'Outro', 'Invalido'),
                   votos = c(1410, 1020, 370, 210))
# Metodo Sison-Glaz (padrao)
print(data.frame(pesq,
                 100 * round( MultinomCI(pesq$votos), 4) ))
```

```
##       cand votos   est lwr.ci upr.ci
## 1        A  1410 46.84  44.95  48.77
## 2        B  1020 33.89  31.99  35.82
## 3    Outro   370 12.29  10.40  14.22
## 4 Invalido   210  6.98   5.08   8.91
```

IC para proporções multinomiais (ii)

```
# Metodo de aproximacao normal ingenua
aprn <- MultinomCI(pesq$votos, method = 'wald')
aprn.df <- data.frame(pesq, 100 * round(aprn, 4) )
print(aprn.df)
```

```
##       cand votos   est lwr.ci upr.ci
## 1        A  1410 46.84  45.06  48.63
## 2        B  1020 33.89  32.20  35.58
## 3    Outro   370 12.29  11.12  13.47
## 4 Invalido   210  6.98   6.07   7.89
```

IC para proporções multinomiais (iii)

```
# E se n = 301?
pesq <- pesq %>% mutate(votos = votos / 10)
# Metodo Sison-Glaz e aproximacao normal (wald) a 95%
sg <- 100*round(MultinomCI(pesq$votos), 4)
colnames(sg) <- c('prop', 'sg.baixo', 'sg.alto')
wald <- 100*round(MultinomCI(pesq$votos, method = 'wald'), 4)
colnames(wald) <- c('prop', 'wald.baixo', 'wald.alto')
# Resultado
(data.frame( pesq, sg, wald[,2:3] ))
```

```
##       cand votos  prop sg.baixo sg.alto wald.baixo wald.alto
## 1        A   141 46.84    41.20   53.06      41.21     52.48
## 2        B   102 33.89    28.24   40.10      28.54     39.23
## 3    Outro    37 12.29     6.64   18.50       8.58     16.00
## 4 Invalido    21  6.98     1.33   13.19       4.10      9.85
```

ICs para médias (i)

O IC para médias de variáveis quantitativas é semelhante ao de proporções:

$$\text{IC} = \text{estimativa de ponto} \pm \text{margem de erro}$$

A estimativa de ponto não viesada da média populacional μ é a média amostral $\bar{y}$.

Pelo TCL, para amostras aleatórias "grandes", a distribuição amostral de $\bar{y}$ é aproximadamente normal.

Portanto, mais uma vez a margem de erro será um **z-score** multiplicado pelo **erro padrão** (com um pequeno detalhe)

ICs para médias (ii)

Vimos que o erro padrão da média amostral é $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$, em que σ é o desvio padrão de y na população.

Como não conhecemos σ, temos que estimá-lo; assim como antes, vamos usar o **desvio padrão amostral s**:

$$se = \frac{s}{\sqrt{n}}$$

Logo, nosso IC será:

$$\bar{y} \pm z \cdot se \rightarrow \bar{y} \pm z \frac{s}{\sqrt{n}}$$

Exercício

Agresti 2018, p. 113

O *GSS* de 2014 coletou informações sobre o número de parceiros sexuais para **129** mulheres entre 23 e 29 anos. Agresti reporta que a média ficou em **6.6**, com desvio padrão de **13.3**. Qual o IC a 95% e 99%?

Exercício

Agresti 2018, p. 113

O *GSS* de 2014 coletou informações sobre o número de parceiros sexuais para **129** mulheres entre 23 e 29 anos. Agresti reporta que a média ficou em **6.6**, com desvio padrão de **13.3**. Qual o IC a 95% e 99%?

```
n <- 129
media <- 6.6
dp <- 13.3
ic <- c(media + qnorm(.025)*dp/sqrt(n),
        media + qnorm(.975)*dp/sqrt(n))
print(ic)
```

```
## [1] 4.304883 8.895117
```

Exercício

Agresti 2018, p. 113

O *GSS* de 2014 coletou informações sobre o número de parceiros sexuais para **129** mulheres entre 23 e 29 anos. Agresti reporta que a média ficou em **6.6**, com desvio padrão de **13.3**. Qual o IC a 95% e 99%?

```
n <- 129
media <- 6.6
dp <- 13.3
ic <- c(media + qnorm(.025)*dp/sqrt(n),
        media + qnorm(.975)*dp/sqrt(n))
print(ic)
```

```
## [1] 4.304883 8.895117
```

...mas na prática não é bem assim, porque temos que estimar σ.

A distribuição t (i)

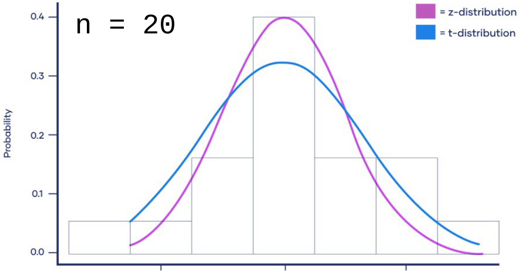
O problema: temos que usar $\hat{\sigma}$ no IC porque não conhecemos σ, o que adiciona erros ao modelo, que podem ser grandes se a amostra for pequena.

Por isso, em vez de usar $Z \sim (0, 1)$ para construir os IC para médias, usamos uma distribuição semelhante, porém mais conservadora em amostras pequenas: a **distribuição t de Student**.

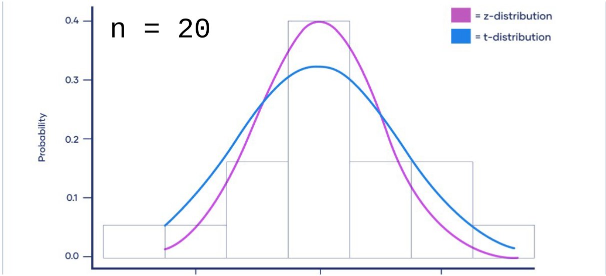
Distribuição t de Student

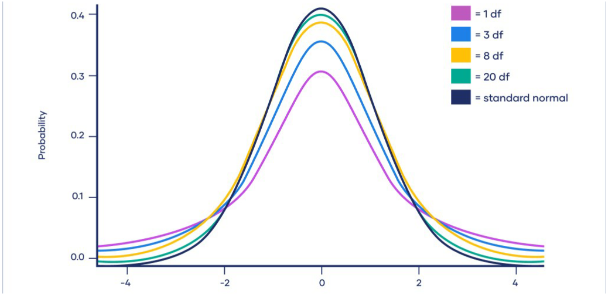
- Em formato de sino, simétrica, com média igual a zero
- O desvio padrão depende dos **graus de liberdade**, convergindo para baixo para 1 quando os g.l. crescem
- Os **graus de liberdade** são obtidos por $gl = n - 1$

A distribuição t (ii)



A distribuição t (ii)





A distribuição t (iii)

Na distribuição normal padrão, os **z-scores** são constantes.

Na distribuição t de Student, os **t-scores** dependem dos graus de liberdade ($n - 1$).

A distribuição t (iii)

Na distribuição normal padrão, os **z-scores** são constantes.

Na distribuição t de Student, os **t-scores** dependem dos graus de liberdade ($n - 1$).

```
##   conf   |z| |t, gl 1| |t, gl 10| |t, gl 100| |t, gl Inf|
## 1 0.90 1.645     6.314      1.812       1.660       1.645
## 2 0.95 1.960    12.706      2.228       1.984       1.960
## 3 0.99 2.576    63.657      3.169       2.626       2.576
```

A distribuição t (iii)

Na distribuição normal padrão, os **z-scores** são constantes.

Na distribuição t de Student, os **t-scores** dependem dos graus de liberdade ($n - 1$).

```
##   conf   |z| |t, gl 1| |t, gl 10| |t, gl 100| |t, gl Inf|
## 1 0.90 1.645     6.314      1.812       1.660       1.645
## 2 0.95 1.960    12.706      2.228       1.984       1.960
## 3 0.99 2.576    63.657      3.169       2.626       2.576
```

Para $n \geq 200$ o t-score já fica muito próximo do z-score...

Para descobrir o t-score no R, use o comando `qt(p, df = gl)`, em que p é a probabilidade acumulada e gl são os graus de liberdade.

Exercício

```
voos <-  flights %>% select(arr_delay) %>%
          filter(!is.na(arr_delay))
descr(voos, stats = c('mean','sd','N.valid'),
      transpose = TRUE, headings = FALSE)
```

```
##
##                    Mean   Std.Dev     N.Valid
## --------------- ------ --------- -----------
##       arr_delay   6.90     44.63   327346.00
```

1. Qual a amplitude do IC com z-score para $n = 40$ a 95%? E com o t-score?
2. Qual a amplitude do IC com z-score para $n = 4000$ a 95%? E com o t-score?
3. Qual deve ser o n para margem de erro ≤ 5 min com 95% de confiança?

Exercício

```
# Tamanho da amostra
n = 40
# Range com z e t a 95% (respectivamente)
2 * qnorm(.975) * sd(voos$arr_delay) / sqrt(n)
2 * qt(.975, df = n - 1) * sd(voos$arr_delay) / sqrt(n)
```

```
## [1] 27.66349
## [1] 28.54884
```

Exercício

```
# Tamanho da amostra
n = 40
# Range com z e t a 95% (respectivamente)
2 * qnorm(.975) * sd(voos$arr_delay) / sqrt(n)
2 * qt(.975, df = n - 1) * sd(voos$arr_delay) / sqrt(n)
```

```
## [1] 27.66349
## [1] 28.54884
```

```
# Repetindo com amostra maior
n = 4000
2 * qnorm(.975) * sd(voos$arr_delay) / sqrt(n)
2 * qt(.975, df = n - 1) * sd(voos$arr_delay) / sqrt(n)
```

```
## [1] 2.766349
## [1] 2.767187
```

Exercício

```
# N para <= 5min de margem de erro
n <- sd(voos$arr_delay)^2 * (qnorm(.975) / 5)^2
print(n)
n <- ceiling(n)
print(n)
```

```
## [1] 306.1075
## [1] 307
```

Exercício

```
# N para <= 5min de margem de erro
n <- sd(voos$arr_delay)^2 * (qnorm(.975) / 5)^2
print(n)
n <- ceiling(n)
print(n)
```

```
## [1] 306.1075
## [1] 307
```

```
# Conferindo
2 * qnorm(.975) * sd(voos$arr_delay) / sqrt(n)
2 * qt(.975, df = n - 1) * sd(voos$arr_delay) / sqrt(n)
```

```
## [1] 9.985454
## [1] 10.0251
```

Robustez

O cálculo do IC da média depende de dois pressupostos:

- Amostragem aleatória
- Distribuição da variável na população é normal

Em estatística, um método é **robusto** com respeito a um pressuposto quando ele tem bom desempenho mesmo que o pressuposto seja violado.

Felizmente, é o caso do IC para média: se $n \geq 15$, o IC baseado na distribuição t funciona bem. Assintoticamente, os problemas desaparecem.

IC por *bootstrap* (i)

Às vezes, não temos informações sobre a distribuição da variável na população, ou queremos estimar um parâmetro cujo comportamento é errático em amostras realistas, ou a fórmula do IC é muito complicada, ou simplesmente temos preguiça.

Nesses casos, se nossa amostra for aleatória, podemos estimar o IC por **bootstrap**:

- O método trata a distribuiçào da variável na nossa amostra como se fosse a distribuição populacional e simula a distribuição amostral.
- Em cada simulação, o método sorteia aleatoriamente, (com reposição), n observações da nossa amostra e calcula a estatística de interesse. Depois repete o procedimento N vezes (para um N bem grande).
- Essa distribuição amostral simulada permite o cálculo de ICs "empíricos".

IC por *bootstrap* (ii)

```
# Dados
voos <-  flights %>%
          select(arr_delay) %>%
            filter(!is.na(arr_delay))

# Media na populacao para referencia
mean(voos$arr_delay)
```

```
## [1] 6.895377
```

IC por *bootstrap* (iii)

```
# Dados
voos_n40 <- voos %>% slice_sample(n = 40) %>% as.vector() %>% un
# IC com t-score
res_t <- t.test(voos_n40)$conf.int[1:2]
# Bootstrap com 10k repeticoes
bs <- boot(voos_n40, function(x,i) mean(x[i]),  R = 10000)
bs.ic <- boot.ci(bs)
res_bs <- rbind(bs.ic$basic[4:5], bs.ic$bca[4:5],
                bs.ic$perc[4:5], bs.ic$normal[2:3])
res <- rbind(res_t, res_bs)
rownames(res) <- c('t-score', 'bs, basic', 'bs, bca',
                   'bs, perc','bs, normal')
colnames(res) <- c('Inferior', 'Superior')
print(res)
```

IC por *bootstrap* (iii)

```
##                Inferior Superior
## t-score       -7.318375 12.61838
## bs, basic     -7.200000 11.82500
## bs, bca       -5.800000 13.35000
## bs, perc      -6.525000 12.50000
## bs, normal    -6.819392 12.12475
```

Próxima aula

Atividade

A atividade #5 será postada no Google Classroom dia 14/11, com prazo para entrega até 21/11

Leituras obrigatórias

Agresti 2018, cap. 6

Leituras optativas

Bussab e Morettin 2010 cap. 120 e 13