Métodos Quantitativos

Aula 06. Fundamentos de probabilidade

Pedro H. G. Ferreira de Souza

pedro.ferreira@ipea.gov.br

Mestrado Profissional em Políticas Públicas e Desenvolvimento

Instituto de Pesquisa Econômica Aplicada (Ipea)

24 out. 2022

- Recapitulação
- Introdução
- Amostragem
- Fundamentos de probabilidade
- Distribuições de probabilidade e variáveis aleatórias
- Distribuições amostrais
- Próxima aula
    - Tarefas para 07/11

Aulas passadas

Aula 01

Metodologia de pesquisa, fundamentos de análises quantitativas

Aula 02

Causalidade, pressupostos do modelo de resultados potenciais

Aulas 03 e 04

Uso de pacotes e funções importação de bases de dados, manipulação de dados no R

Aula 05

Mais funções para o R, estatísticas descritivas simples, visualização de dados

Recapitulação

Introdução

Amostragem

Fundamentos de probabilidade

Distribuições de probabilidade e variáveis aleatórias

Distribuições amostrais

Próxima aula

- Tarefas para 07/11

Sobre amostras e populações

Até aqui, examinamos apenas os dados coletados - nossas **amostras** - sem preocupação em extrapolar os resultados para um **universo** ou **população** de interesse.

Na prática, contudo, o que nos interessa é quase sempre o **parâmetro populacional**, não a **estatística amostral** $\rightarrow$ problema de **inferência estatística**

Perguntas

- Como usar a amostra para produzir uma *boa* **estimativa** do parâmetro populacional?
- Como quantificar a **incerteza** relativa à nossa estimativa?

De onde vêm os erros?

O que pode dar errado quando você produz uma estimativa a partir de uma amostra?

Viés amostral

A amostra não representa adequadamente a população de interesse, com sub/sobre-representação de subgrupos relevantes

Viés de resposta ou de captação

Instrumento de coleta não registra corretamente as características e/ou opiniões das unidades

Viés de não resposta

Subgrupo não aleatório das unidades amostradas não é encontrada, se recusa a participar ou não tem todas suas características registradas

Um caso clássico

Em 1936, a revista americana *Literary Digest* enviou 10m de cartões postais, obtendo mais de 2m de respostas, e proclamou que **Alfred Landon** seria eleito presidente com **57%** do voto popular e **370** votos no colégio eleitoral...

Um caso clássico

Em 1936, a revista americana *Literary Digest* enviou 10m de cartões postais, obtendo mais de 2m de respostas, e proclamou que **Alfred Landon** seria eleito presidente com **57%** do voto popular e **370** votos no colégio eleitoral...

...mas **Franklin Roosevelt** foi eleito com **61%** do voto popular e **523** votos no colégio eleitoral, um dos maiores massacres eleitorais da história americana.

Quem acertou foi George Gallup, que previu **56%** do voto popular e **481** votos no colégio eleitoral para o Roosevelt com uma amostra de *apenas* 50 mil.

(Não houve CPI, ninguém foi preso nem perseguido, mas a revista perdeu credibilidade e faliu um ano e meio depois. O Gallup ficou rico.)

Outra fórmula para variância

Na aula passada vimos a estatística amostral:

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

Para os parâmetros, usaremos hoje:

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Obs: nos slides, uso a notação de um jeito mais frouxo para facilitar.

Pacotes e dados de hoje

```
library(tidyverse)
library(summarytools)
library(nycflights13)
voos.df <- flights
```

Problema central

A qualidade das inferências estatísticas depende de quão bem a amostra representa a população. Logo, é necessário:

- Definir a população de interesse
- Utilizar um mecanismo de seleção da amostra que "garanta" a representatividade
- Escolher um tamanho adequado para a amostra

→ Quanto mais informações tivermos sobre a população, melhor.

→ Objetivo é minimizar o risco de **viés de seleção**

Aleatorização

Mais uma vez, o **sorteio aleatório** de casos oferece a melhor proteção contra **viés de seleção**. Por isso, amostras são classificadas em:

Amostras aleatórias ou probabilísticas

- Cada unidade da população tem uma probabilidade conhecida de ser sorteada devido à utilização de alguma foram de seleção aleatória

Amostras não aleatórias ou não probabilísticas

- Não sabemos as probabilidades de inclusão de cada unidade e, portanto, não podemos usar teoria probabilística para quantificar o viés e a incerteza das nossas estimativas

Amostras aleatórias

Amostra aleatória simples (AAS)

Todas as unidades da população têm a mesma probabilidade de serem sorteadas, ou seja, cada subconjunto com n unidades tem a mesma probabilidade de ser sorteado que qualquer outro subconjunto de tamanho n.

Amostras aleatórias complexas

Estratificação $\rightarrow$ para garantir representatividade, particionamos a população de forma exaustiva e extraímos AAS de cada partição

Conglomerados $\rightarrow$ particionamos a população de forma exaustiva, extraímos AAS de n partições e coletamos informações de todas as unidades da partição

(e vários outros tipos)

Selecionando uma amostra aleatória simples

Basta enumerar todas as unidades de uma população de tamanho N e usar um gerador de números aleatórios para sortear n unidades.

Amostragem sem reposição

- Cada unidade só pode ser sorteada uma vez
- Logo, não há independência entre os sorteios e a covariância entre valores sorteados não é zero
- A probabilidade de inclusão para uma unidade é r/N

Amostragem com reposição

- Cada unidade pode ser sorteada mais de uma vez
- Independência entre sorteios e a covariância é zero $\rightarrow$ mais fácil de lidar matematicamente, por isso é preferido
- A probabilidade de inclusão para uma unidade é $1 - (1 - \frac{1}{N})^n$

AAS no R com slice_sample()

Para sortear *n* linhas:

```
sem_reposicao.df <- voos.df %>%
                      slice_sample(n = 1000, replace = FALSE)
com_reposicao.df <- voos.df %>%
                      slice_sample(n = 1000, replace = TRUE)
```

Para sortear uma proporção *p* das linhas:

```
sem_reposicao.df <- voos.df %>%
                      slice_sample(prop = 0.01, replace = FALSE)
com_reposicao.df <- voos.df %>%
                      slice_sample(prop = 0.01, replace = TRUE)
```

Cuidado!

Não confundam **amostragem aleatória** com **tratamento aleatório**

- Amostragem aleatória $\rightarrow$ como selecionamos as unidades que participarão do estudo
- Tratamento aleatório $\rightarrow$ como alocamos o tratamento em estudos experimentais de causalidade

Ambos são fonte de **incerteza** sobre nossas estimativas, mas podem ou não ser combinados.

O que é probabilidade?

Interpretação clássica

$$P(A) = \frac{N_A}{N} = \frac{eventos}{espacoamostral}$$

Frequências relativas

$$P(A) \approx \frac{n_A}{n}, \qquad n \to \infty$$

(Há também interpretações subjetivas ou bayesianas)

Espaço amostral

É o conjunto de todos os possíveis resultados aleatórios

- Pode ser *contável* ou não, infinito ou não
- Qualquer subconjunto A de Ω é um *evento*

Exemplo: qual o espaço amostral para o lançamento de dois dados?

$$
\begin{aligned}
\Omega = \{&(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\
&(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\
&(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\
&(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\
&(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\
&(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}
\end{aligned}
$$

Exemplos

Para o lançamento de dois dados, qual a probabilidade de...

- Ambos caírem no número 6?
- Ambos caírem em números ímpares?
- A soma de ambos ser maior do que 9?
- O produto de ambos ser menor do que 6?

Exemplos

Para o lançamento de dois dados, qual a probabilidade de...

- Ambos caírem no número 6?
- Ambos caírem em números ímpares?
- A soma de ambos ser maior do que 9?
- O produto de ambos ser menor do que 6?

Respostas:

- $P(6,6) = \frac{1}{36} \approx 3\%$
- $P(I,I) = \frac{9}{36} = 25\%$
- $P(\text{soma} > 9) = \frac{6}{36} \approx 17\%$
- $P(\text{produto} < 6) = \frac{10}{36} \approx 28\%$

Regras básicas de probabilidade

$$0 \leq P(A) \leq 1, \quad A \subseteq \Omega$$

$$P(\Omega) = 1$$

$$P(A^c) = 1 - P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

Paradoxo dos aniversários

Nossa turma tem 29 alunos. Qual a probabilidade de que pelo menos dois de vocês tenham a mesma data de aniversário?

Paradoxo dos aniversários

Nossa turma tem 29 alunos. Qual a probabilidade de que pelo menos dois de vocês tenham a mesma data de aniversário?

Probabilidade de *ninguém* compartilhar aniversário:

$$P(A^c) = \frac{365}{365}\frac{364}{365}\frac{363}{365}\cdots\frac{337}{365} = \frac{365\cdot 364\cdot 363\cdot\ldots\cdot 337}{365^{29}} \approx 31.9\%$$

Paradoxo dos aniversários

Nossa turma tem 29 alunos. Qual a probabilidade de que pelo menos dois de vocês tenham a mesma data de aniversário?

Probabilidade de *ninguém* compartilhar aniversário:

$$P(A^c) = \frac{365}{365}\frac{364}{365}\frac{363}{365}\cdots\frac{337}{365} = \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot 337}{365^{29}} \approx 31.9\%$$

Logo, probabilidade de que alguém compartilhe:

$$P(A) = 1 - P(A^c) = 1 - 0.319 \approx 68.1\%$$

Paradoxo dos aniversários

Usando o R para calcular:

```
alunos <- data.frame(id = seq(1:29))
alunos <- alunos %>% mutate(dias_unicos = 365 - id + 1)

num_eventos <- prod(alunos$dias_unicos)
espaco_amostral <- 365^nrow(alunos)

print(1 - num_eventos / espaco_amostral) # Prob(A)

## [1] 0.6809685
```

(Para quem se interessar, há uma fórmula geral)

Probabilidade condicional

Para dois eventos quaisquer, A e B, sendo $P(B) > 0$, definimos a probabilidade condicional de A dado B como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

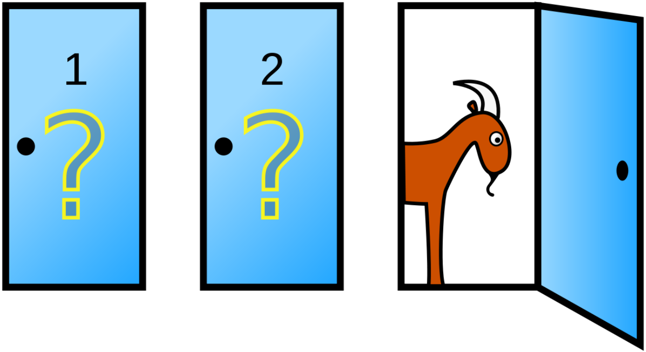
Exemplo: um casal tem dois filhos. Sabemos que um deles é homem. Qual a probabilidade do outro também ser homem?

$$\Omega = \{(H, H), (H, M), (M, H)\}$$

$$P(A|B) = \frac{1}{3}$$

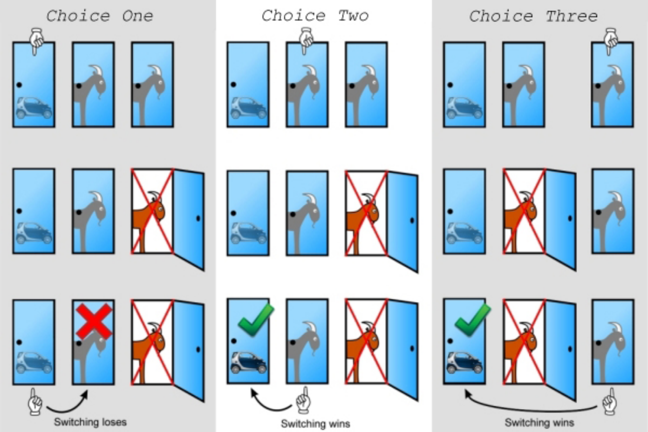
Exercício

Problema de Monty Hall ou *da Porta da Esperança*



Trocar de porta ou não? Quais as probabilidades?

Solução visual



Vence se trocar $\rightarrow \frac{1}{3}0 + \frac{1}{3}1 + \frac{1}{3}1 = \frac{2}{3} \approx 66\%$

Vence se *não* trocar $\rightarrow \frac{1}{3}1 + \frac{1}{3}0 + \frac{1}{3}0 \approx 33\%$

Solução por probabilidades condicionais

Teorema de Bayes

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Solução por probabilidades condicionais

Teorema de Bayes

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Seja A o evento em que o carro está na porta 1 e B o evento em que ele abre a porta 2. Como selecionamos a porta 1, $P(A|B)$ é a nossa probabilidade de vitória se **não** trocarmos.

Temos $P(A) = 1/3$ e $P(B|A) = 1/2$. Só precisamos de $P(B)$, depois é fácil:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{(1/3)(1/2)}{(1/2 + 0 + 1)/3} = \frac{1/6}{1/2} = \frac{1}{3} \approx 33\%$$

Logo, se trocarmos, a probabilidade de vitória é $1 - \frac{1}{3} \approx 66\%$.

Independência

Vimos que $P(A \cap B) = P(A) \cdot P(B|A)$ e, de modo equivalente, $P(A \cap B) = P(B) \cdot P(A|B)$, certo? Mas se os eventos forem **independentes** o cálculo fica mais simples.

Definição de independência

$$P(A \cap B) = P(A) \cdot P(B)$$

Ou seja: $P(A|B) = P(A)$ e $P(B|A) = P(B)$.

Esse é um conceito **crucial** para inferência estatística porque facilita muitos cálculos.

Introdução

Cada evento ou observação de uma **variável aleatória** gera resultados variáveis que podem ser resumidos em probabilidades.

Ou seja, variáveis aleatórias possuem **distribuições de probabilidade** que associam valores à sua probabilidade de ocorrência.

Variáveis aleatórias podem ser **discretas** ou **contínuas**.

Variáveis aleatórias discretas

Uma VA discreta é uma variável que assume um número finito ou "contável" de valores.

Seja $p(x) = p(X = x)$ a probabilidade de um resultado x para a variável X. Ou seja, $p(x)$ é a **função de distribuição de probabilidade** da variável X. Então:

$$0 \leq p(x) \leq 1$$

$$\sum p(x) = 1$$

A **função de distribuição acumulada** é dada por $P(x) = P(X \leq x)$

Exemplo com dado não viciado

x	p(x)	P(x)
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	6/6

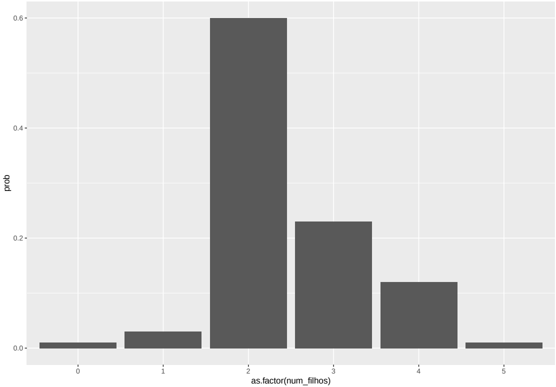
Exemplo de Agresti 2018, p. 70

```
agresti_tab41.df <-
  data.frame(num_filhos = c(0, 1, 2, 3, 4, 5),
             prob = c(.01, .03, .60, .23, .12, .01))
print(agresti_tab41.df, row.names = FALSE)
```

```
##  num_filhos prob
##           0 0.01
##           1 0.03
##           2 0.60
##           3 0.23
##           4 0.12
##           5 0.01
```

Exemplo de Agresti 2018, p. 70

```
qplot(data = agresti_tab41.df,
      x = as.factor(num_filhos),
      y = prob, geom = 'col')
```



Média, variância e desvio-padrão de uma VA discreta

Suponha uma VA discreta X com valores finitos $x_1, x_2, \ldots x_N$, cada um com probabilidade p_i:

Valor médio ou esperança matemática

$$E(X) = \mu = \sum_{i=1}^{N} x_i p_i$$

Variância

$$Var(X) = E[(X - \mu)^2] = \sum_{i=1}^{N} p_i (x_i - \mu)^2$$

Lembrem-se que $sd = \sqrt{var}$

Exemplo de Agresti 2018, p. 70

```
with(agresti_tab41.df, descr(num_filhos,
                             stats = c('mean','sd'),
                             weights = 100*prob))
```

```
## Weighted Descriptive Statistics
## agresti_tab41.df$num_filhos
## Weights: 100 * prob
## N: 6
##
##                   num_filhos
## ------------- ------------
##          Mean         2.45
##       Std.Dev         0.82
```

Distribuição uniforme discreta

$X \sim U(a, b)$, no caso de variáveis discretas, significa que cada valor inteiro entre a e b ocorre com a mesma probabilidade, dada por:

$$p(x) = \frac{1}{b - a + 1}, \qquad x = a, a + 1, \ldots, b$$

O valor esperado a variância são:

$$E(X) = \frac{1}{b - a + 1} \sum_{i=a}^{b} i = \frac{a + b}{2}$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{(b - a + 1)^2 + 1}{12}$$

Distribuição uniforme discreta

```r
# Cria uma sequencia uniforme discreta
dunifd <- seq(from = 42, to = 82, by = 1)
# Objetos com o numero de valores, minimo e maximo
n <- length(dunifd);  a <- min(dunifd); b <- max(dunifd)
# Valor esperado (media)
mean(dunifd)
(a + b) / 2
# Variancia (observem a discrepancia!)
var(dunifd)
(n^2 + 1) / 12
```

```
## [1] 62
## [1] 62
## [1] 143.5
## [1] 140.1667
```

Distribuição de Bernoulli

Uma variável aleatória de Bernoulli, $X \sim Ber(p)$, assume apenas os valores 0 e 1:

$$p(x = 0) = 1 - p, \quad p(x = 1) = p$$

O valor esperado e a variância são:

$$E[X] = \sum x_i p_i = (1 - p) \cdot 0 + p \cdot 1 = p$$

$$Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1 - p)$$

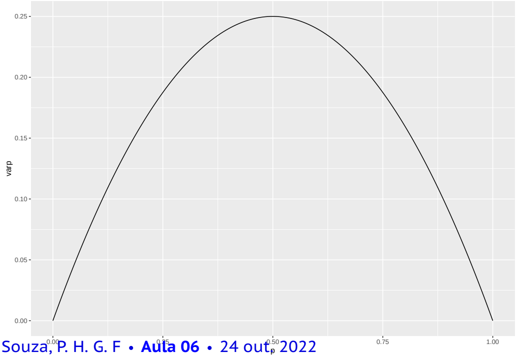
Distribuição de Bernoulli

A variância de uma VA Bernoulli depende apenas do parâmetro p, que estabelece o percentual de sucesso. Qual p maximiza a variância?

Distribuição de Bernoulli

A variância de uma VA Bernoulli depende apenas do parâmetro p, que estabelece o percentual de sucesso. Qual p maximiza a variância?

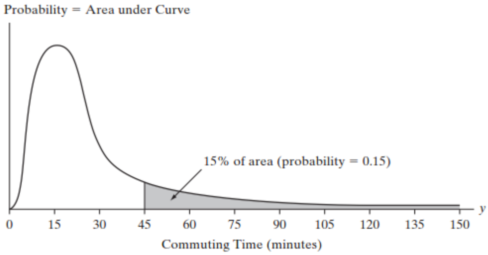
```
bernvar.df <- data.frame(p = seq(from = 0, to = 1, length.out = 1001))
bernvar.df <- bernvar.df %>% mutate(varp = p * (1 - p))
qplot(data = bernvar.df, x = p, y = varp, geom = 'line')
```



Variáveis aleatórias contínuas

Uma VA contínua pode assumir infinitos valores, sendo impossível atribuir probabilidades específicas a cada valor. Nesse caso, atribuímos probabilidade a intervalos, e a probabilidade do intervalo que contém todos os valores possíveis é igual a 1.

Em vez de um histograma, a representação visual é uma curva contínua de densdiade, e as probabilidades representam determinada área sob a curva.



Distribuição uniforme contínua

Se $X \sim U(a, b)$:

$$PDF \rightarrow f(x; a, b) = \frac{1}{b - a}, \quad para a \leq x \leq b$$

O valor esperado não muda em relação a antes, mas a variância sim:

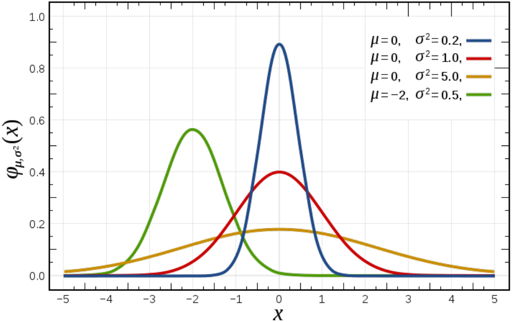
$$E[X] = \int_a^b xf(x)dx = \int_a^b \frac{x}{b - a}dx = \frac{a + b}{2}$$

$$Var(x) = \frac{(b - a)^2}{12}$$

Distribuição normal $N(\mu, \sigma^2)$

A distribuição normal é simétrica, com formato de sino, caracterizada pela média μ e pelo variância σ^2. Sua densidade é dada por:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$



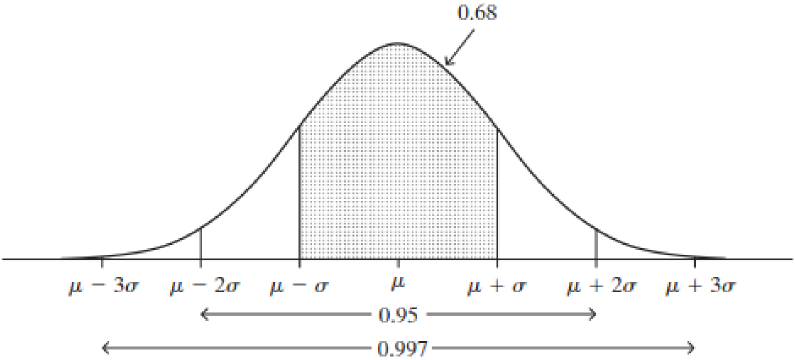
Exemplos da vida real

Muitos fenômenos observáveis têm distribuição (aproximadamente) normal, como:

- Peso
- Altura
- Pressão sanguínea
- Temperatura
- Desempenho em provas
- Tamanhos de sapatos
- etc

Distribuição normal $N(\mu, \sigma^2)$

Propriedades úteis dessa distribuição de probabilidade:



Função pnorm() no R

```
media <- 100
dp <- 20
ate_menos1dp <- pnorm( media - dp, mean = media, sd = dp)
ate_1dp <- pnorm( media + dp, mean = media, sd = dp)
ate_1dp - ate_menos1dp
```

```
## [1] 0.6826895
```

Função pnorm() no R

```
media <- 100
dp <- 20
ate_menos1dp <- pnorm( media - dp, mean = media, sd = dp)
ate_1dp <- pnorm( media + dp, mean = media, sd = dp)
ate_1dp - ate_menos1dp
```

```
## [1] 0.6826895
```

```
media <- pi
dp <- 0.13
ate_menos2dp <- pnorm( media - 2*dp, mean = media, sd = dp)
ate_2dp <- pnorm( media + 2*dp, mean = media, sd = dp)
ate_2dp - ate_menos2dp
```

```
## [1] 0.9544997
```

Função qnorm() no R

```
qnorm(.025, mean = 0, sd = 1)
qnorm(.975, mean = 0, sd = 1)
qnorm(.500, mean = 0, sd = 1)
```

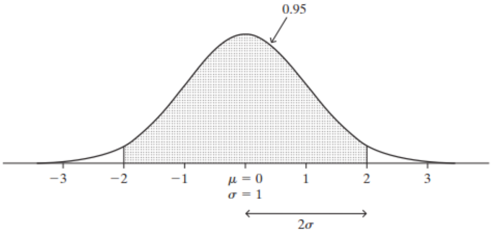
```
## [1] -1.959964
## [1] 1.959964
## [1] 0
```

Distribuição normal padronizada $N(0, 1)$

Como a área sob a curva é constante em múltiplos de σ, podemos **padronizar** uma curva normal para obter os **z-score** dos valores:

$$z = \frac{x - \mu}{\sigma}$$

Logo, $Z \sim N(0, 1)$



Exercício

Suponha que a altura de homens adulto segue uma distribuição normal com média de 1.80m e desvio-padrão de 10cm:

$$X \sim N(1.80, 0.01)$$

1. Qual a probabilidade de que um homem sorteado aleatoriamente tenha menos de 1.80?
2. Qual a probabilidade de que ele tenha entre 1.60m e 1.80m?
3. Qual a probabilidade de que ele tenha mais de 2m?
4. Qual a altura mínima para estar no 1% mais alto?

Exercício

```
# 1. Probabilidade de ter ate 1.8 = Pr(X < 1.80)
pnorm(1.80, mean = 1.8, sd = .1)
# 2. Probabilidade de ter mais de 2m = 1 - Pr(X < 2)
1 - pnorm(2, mean = 1.8, sd = .1)
```

```
## [1] 0.5
## [1] 0.02275013
```

Exercício

```
# 1. Probabilidade de ter ate 1.8 = Pr(X < 1.80)
pnorm(1.80, mean = 1.8, sd = .1)
# 2. Probabilidade de ter mais de 2m = 1 - Pr(X < 2)
1 - pnorm(2, mean = 1.8, sd = .1)
```

```
## [1] 0.5
## [1] 0.02275013
```

```
# 3. Probabilidade de ter entre 1.5 e 1.6 =
# Pr(X < 1.8) - Pr(X < 1.6)
pnorm(1.8,mean=1.8,sd=.1) - pnorm(1.6,mean=1.8,sd=.1)
# 4. Altura minima para o top 1% = Pr(X < y) = .99
qnorm(0.99, mean = 1.8, sd = .1)
```

```
## [1] 0.4772499
## [1] 2.032635
```

Parâmetros e estatísticas

```
# Qual a distancia media percorrida pelos voos que chegam a NY?
voos.df %>% descr(var = distance, stats = c('n.valid', 'mean'),
                  transpose = TRUE, headings = FALSE)
```

```
##
##                     N.Valid      Mean
## -------------- ----------- ---------
##       distance   336776.00   1039.91
```

Parâmetros e estatísticas

```
# Qual a distancia media percorrida pelos voos que chegam a NY?
voos.df %>% descr(var = distance, stats = c('n.valid', 'mean'),
                  transpose = TRUE, headings = FALSE)

##
##                     N.Valid      Mean
## -------------- ----------- ---------
##       distance   336776.00   1039.91

# Em uma AAS com n = 100, que valor medio obtemos?
slice_sample(voos.df, n = 100) %>%
  descr(var = distance, stats = c('n.valid', 'mean'),
        transpose = TRUE, headings = FALSE)

##
##                   N.Valid     Mean
## -------------- --------- --------
##       distance    100.00   970.55
```

Parâmetros e estatísticas

```r
# E se tirarmos zilhoes de amostras?
dist_obs <- as.vector(voos.df$distance)
amostras <- replicate(5000, mean( sample(dist_obs, size = 100) ) )
amostras <- data.frame(media = amostras)
```

Parâmetros e estatísticas

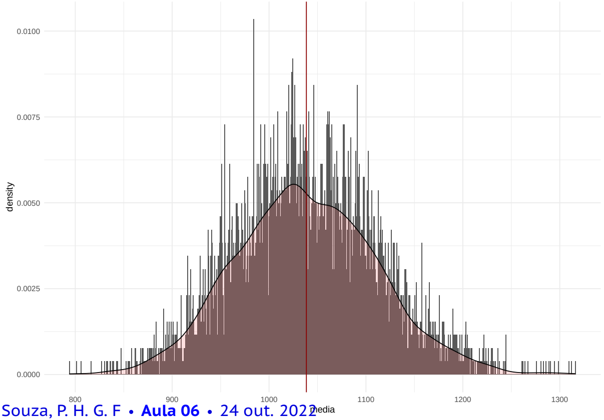
```
# E se tirarmos zilhoes de amostras?
dist_obs <- as.vector(voos.df$distance)
amostras <- replicate(5000, mean( sample(dist_obs, size = 100) ) )
amostras <- data.frame(media = amostras)
```

```
amostras %>% descr(stats = c('n.valid', 'mean', 'q1', 'med', 'q3'),
                   headings = FALSE, transpose = TRUE)
```

```
##
##                 N.Valid      Mean       Q1    Median        Q3
## ----------- --------- --------- -------- --------- ---------
##       media   5000.00   1038.51   988.62   1035.67   1087.78
```

Parâmetros e estatísticas

```
ggplot(amostras, aes(x = media)) +
  geom_histogram(aes(y=..density..), bins = 1000) +
  geom_density(alpha = .2, fill = 'indianred1') +
  geom_vline(aes(xintercept = mean(media)), color = 'darkred') +
  theme_minimal()
```



Distribuição amostral

Definição

A distribuição amostral de uma estatística é a distribuição de probabilidade que especifica as probabilidades para os valores que a estatística pode assumir.

A distribuição amostral diz respeito à variabilidade da estatística de interesse em diferentes amostras de mesmo tamanho.

Ponto central: nossa amostra observada é sempre apenas uma entre muitas possíveis.

Logo, uma estatística amostral é uma **variável aleatória** porque se baseia em amostras aleatórias de uma população. Por isso, ela possui uma distribuição de probabilidade, que é a **distribuição amostral**.

Erro padrão

Definição

O **erro padrão** é o desvio padrão da distribuição amostral da estatística.

- Imagine que sorteamos zilhões de amostras com tamanho n e calculamos uma estatística (por exemplo, a média). Se criarmos uma variável com o valor da estatística em cada amostra, o erro padrão é o desvio padrão dessa variável.

```
sd(amostras$media)
```

```
## [1] 72.89981
```

O erro padrão de uma estatística nos diz a variabilidade dela em diferentes amostras de mesmo tamanho.

Distribuição amostral da média

Teorema Central do Limite

Considere uma amostra aleatória de tamanho n de uma variável com média populacional igual a μ e desvio padrão igual a σ.

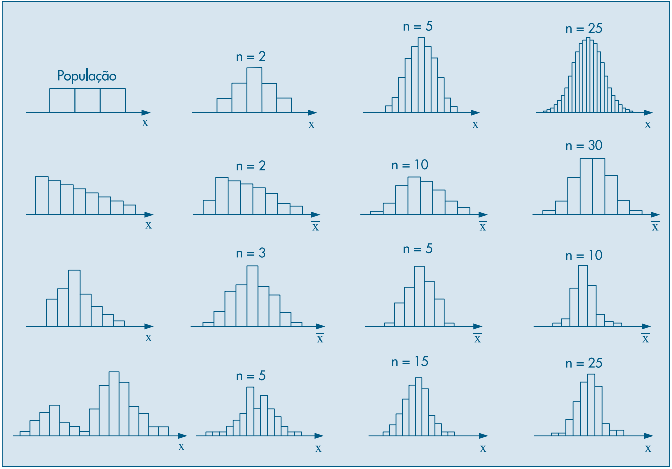
A distribuição amostral da média amostral $\bar{x}$ tem (aproximadamente) a forma de uma distribuição normal com parâmetros:

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

- O TCL vale para todas as VA, independentemente da sua distribuição: a única coisa que muda é que distribuições mais assimétricas exigem amostras maiores ($n \geq 30$) para aproximar melhor uma distribuição normal

Distribuição amostral da média

Figura 10.5: Histogramas correspondentes às distribuições amostrais de $\overline{X}$ para amostras extraídas de algumas populações.



Distribuição amostral da média

Vamos testar empiricamente:

```
# Obtendo o erro padrao a partir
# do parametro conhecido
sd(voos.df$distance) / sqrt(100)
```

```
## [1] 73.3233
```

```
# Estimando o erro padrao pelo o SD
# das zilhoes de amostras
sd(amostras$media)
```

```
## [1] 72.89981
```

Distribuição amostral da média

Por que isso importa?

Como vimos, já sabemos as probabilidades associadas a uma distribuição normal: a estatística calculada a partir de uma única amostra tem probabilidade de...

- $\sim$ 68% de ficar entre $(\mu - \sigma, \mu + \sigma)$
- $\sim$ 95% de ficar entre $(\mu - 2\sigma, \mu + 2\sigma)$
- $\sim$ 99% de ficar entre $(\mu - 3\sigma, \mu + 3\sigma)$

Nas próximas aulas, vamos usar isso para quantificar a incerteza das nossas estimativas.

Distribuição amostral de outras estatísticas

O TCL pode ser estendido para várias outras estatísticas, mas nem sempre a conclusão é tão geral e não necessariamente a distribuição amostral é normal. Por exemplo:

Variância

Distribuiçào amostral é um múltiplo da distribuição χ^2 (qui-quadrado) quando na população a variável tem distribuição normal.

Mediana

Distribuição amostral assintoticamente normal com média centrada na mediana e $var(mediana) = \pi\sigma^2/2n$.

O tamanho das amostras e o erro padrão

Como dito, o **erro padrão** nos diz a variabilidade amostral da estatística.

No caso da **média amostral**, o erro padrão é dado por $\sigma/\sqrt{n}$, em que n é o tamanho da amostra. Ou seja, para uma variável x qualquer, a variabilidade das nossas estimativas de $\bar{x}$ depende de:

- Desvio padrão de x na população: quanto maior, maior a variabilidade de $\bar{x}$ (e vice-versa)
- O tamanho n da nossa amostra: quanto maior a amostra, menor a variabilidade de $\bar{x}$ (e vice-versa)

Ou seja, o erro padrão diminui conforme n aumenta, mas a relação **não é linear**.

Exemplo

Suponha uma variável $X \sim N(1000, 10, 000)$, ou seja, com valor esperado $\mu = 1000$ e desvio padrão $\sigma = 100$.

Queremos estimar $\bar{x}$. A tabela abaixo calcula os parâmetros da distribuição amostral de $\bar{x}$ para amostras de diferentes tamanhos:

n	$\mu_{\bar{x}}$	$\sigma_{\bar{x}}$
10	1,000	31.6
50	1,000	14.1
100	1,000	10.0
500	1,000	4.5
1,000	1,000	3.2
10,000	1,000	1.0
100,000	1,000	0.3

Exemplo no R

Se alguém quiser ver para crer, podemos simular essas distribuições. Por exemplo, vamos simular a distribuição amostral com $n = 100$:

```
# Simulando para n = 100
sim_n10 <- replicate(5000, rnorm(n = 100, mean=1000, sd=100))
medias_n10 <- colMeans(sim_n10)
mean(medias_n10)
```

```
## [1] 1000.101
```

```
sd(medias_n10)
```

```
## [1] 10.04133
```

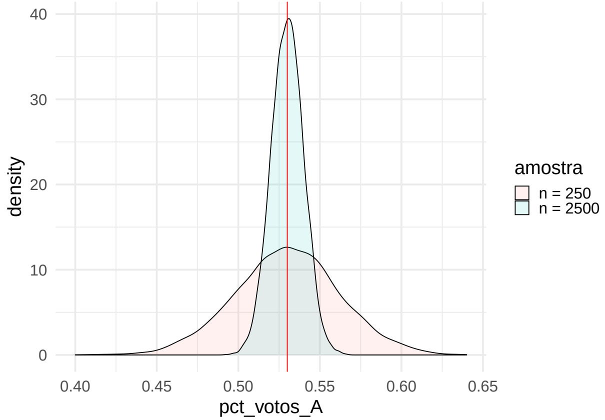
Exemplo eleitoral

Pense no segundo turno de uma eleição em que 53% votam no candidato A e 47% votam no candidato B. Como é a distribuição amostral para $n = 250$ e $n = 2,000$?

```
# Amostras com n = 250
sim_n250  <- replicate( 5000, rbernoulli(250, p = 0.53))
sim_n250.df <- data.frame(amostra = 'n = 250',
                          pct_votos_A = colMeans(sim_n250))
# Amostra com n = 2500
sim_n2500 <- replicate(5000, rbernoulli(2500, p = 0.53))
sim_n2500.df <- data.frame(amostra = 'n = 2500',
                           pct_votos_A = colMeans(sim_n2500))
# Junta os dois data frames
sim_n.df <- rbind(sim_n250.df, sim_n2500.df)
```

Exemplo eleitoral

```
ggplot(sim_n.df, aes(x = pct_votos_A, fill = amostra)) +
  theme_minimal(base_size = 22) + geom_density(alpha = .1) +
  geom_vline(xintercept = 0.53, color = 'red')
```



Exemplo eleitoral

```
sim_n.df %>%
  group_by(amostra) %>%
    summarise(n = n(),
              media = mean(pct_votos_A),
              p02.5 = quantile(pct_votos_A, prob = 0.025),
              p50 = quantile(pct_votos_A, prob = 0.500),
              p97.5 = quantile(pct_votos_A, prob = 0.975))
```

```
## # A tibble: 2 x 6
##   amostra      n media p02.5   p50 p97.5
##   <chr>    <int> <dbl> <dbl> <dbl> <dbl>
## 1 n = 250   5000 0.530 0.464 0.532 0.592
## 2 n = 2500  5000 0.530 0.511 0.53  0.549
```

Determinação do tamanho da amostra

Queremos estimar a média populacional μ com base na média amostral $\bar{x}$ para uma amostra de tamanho n de modo que $P(-\epsilon \leq \bar{x} - \mu \leq \epsilon) \geq \gamma$

- ϵ é a **margem de erro** que estamos dispostos a tolerar
- γ é o **grau de confiança** desejado para estar dentro da margem

A distribuição amostral de $\bar{x}$ é $N(\mu, \sigma^2/n)$. Logo, a de $\bar{x} - \mu$ é $N(0, \sigma^2/n)$. Para a normal padrão, basta dividir por $\sigma/\sqrt(n)$.

$$P(-\epsilon \leq \bar{x} - \mu \leq \epsilon) = P(-\frac{\sqrt{n}\epsilon}{\sigma} \leq Z \leq \frac{\sqrt{n}\epsilon}{\sigma}) \approx \gamma$$

Nós escolhemos γ, então obtemos z_x da $N(0,1)$, tal que $P(-z_x \leq Z \leq z_x) = \gamma$:

$$\frac{\sqrt{n}\epsilon}{\sigma} = z_x \rightarrow n = \frac{\sigma^2 z_x^2}{\epsilon^2}$$

Continuando o exemplo eleitoral

Margem de erro escolhida: 2 pontos percentuais ($\epsilon = 0.02$)

Grau de confiança: 95% ($\gamma = 0.95$)

Ou seja, se fizermos várias pesquisas, nosso resultado vai estar dentro da margem de erro em 95% delas.

Sabemos que a variância de uma VA Bernoulli atinge o valor máximo quando $p = 0.5$, de modo que $var(x) = p(1 - p) = 0.25$. Logo:

$$n = \frac{\sigma^2 z_x^2}{\epsilon^2} = \frac{0.25 \cdot 1.96^2}{0.02^2} \approx 2401$$

Obs: $z_x = 1.96$ porque, como vimos, 95% da área sob a distribuição normal padrão está entre $\mu - 1.96$ e $\mu + 1.96$.

Continuando o exemplo eleitoral

Para γ = 95%:

ϵ	n
0.03	1,067
0.02	2,401
0.01	9,604
0.005	38,416

Para ϵ = 0.02:

γ	n
90%	1,691
95%	2,401
99%	4,147
99.9%	6,768

Recapitulação

Introdução

Amostragem

Fundamentos de probabilidade

Distribuições de probabilidade e variáveis aleatórias

Distribuições amostrais

Próxima aula

- Tarefas para 07/11

Próxima aula

Leituras obrigatórias

Agresti 2018, cap. 5

Leituras optativas

Bussab e Morettin 2010 cap. 10 e 11