Análise Quantitativa de Dados em Linguística

valor de *p*

Ronaldo Lima Jr.

ronaldojr@letras.ufc.br

ronaldolimajr.github.io

Universidade Federal do Ceará

Roteiro

População e Amostra

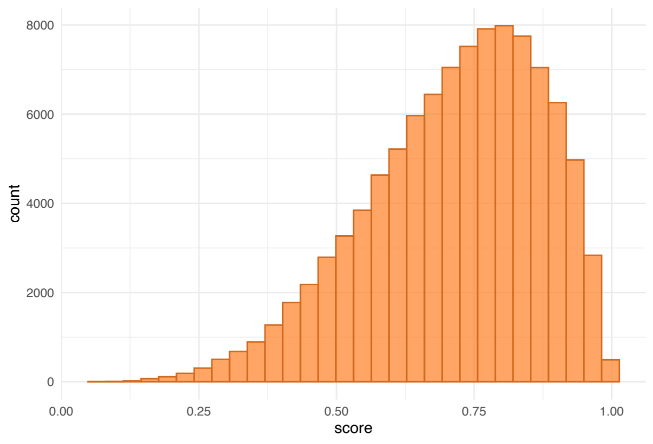
População e Amostra

O grande problema do pesquisador, e consequente objetivo da estatística inferencial, é inferir parâmetros (desconhecidos) de uma população com base nos dados (conhecidos) de uma amostra.

- Podemos simular uma população no R (conhecendo seus parâmetros reais) e extrair amostras aleatórias para ver como os dados se comportam em relação aos parâmetros
- Simular população de 100 mil aluno/as e suas notas em uma prova:

```
population = rbeta(100000, 5, 2)
summary(population)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.06605 0.61107 0.73605 0.71469 0.83866 0.99933

sd(population)
[1] 0.1594129
```

Simulando uma população



Lembre-se: $\mu = 71,5 \ \sigma = 16$

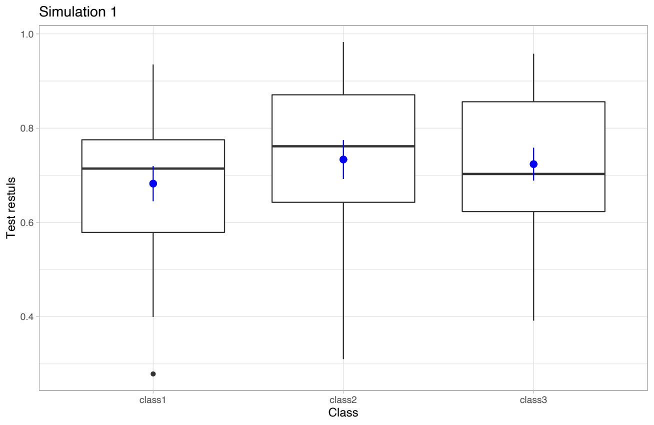
Simulando amostras

Podemos extrair 3 amostras aleatórias dessa população, cada uma com 20 aluno/as, simulando 3 turmas diferentes de aluno/as vindo/as da mesma população, e verificar as médias ($\overline{X}$) e desvios-padrão (S) das amostras:

```
sample1 = sample(x = population, size = 20)
sample2 = sample(x = population, size = 20)
sample3 = sample(x = population, size = 20)

# Criar um tibble (data frame) com os dados das 3 turmas simuladas (samples)
sample.data = tibble(class1 = sample1,
                     class2 = sample2,
                     class3 = sample3) %>%
  gather("class1", "class2", "class3", key = class, value = test)

# Verificar média e desvio-padrão dos dados das turmas
sample.data %>%
  group_by(class) %>%
  summarize(Test.mean = mean(test),
            Test.SD = sd(test))
```

Simulando amostras

```
   class  Test.mean Test.SD
1 class1      0.682   0.167
2 class2      0.734   0.185
3 class3      0.724   0.156
```

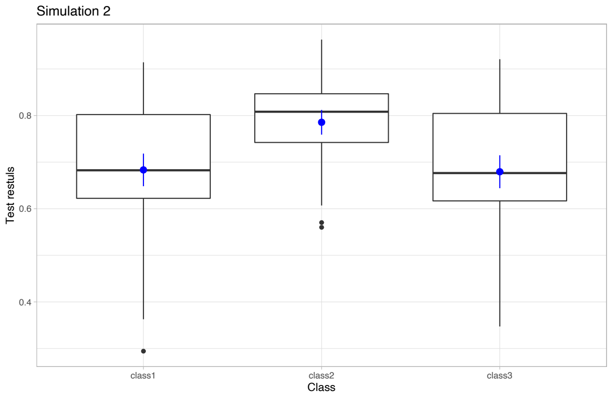


Lembrando: $\mu = 71,5$ $\sigma = 16$

Simulando amostras

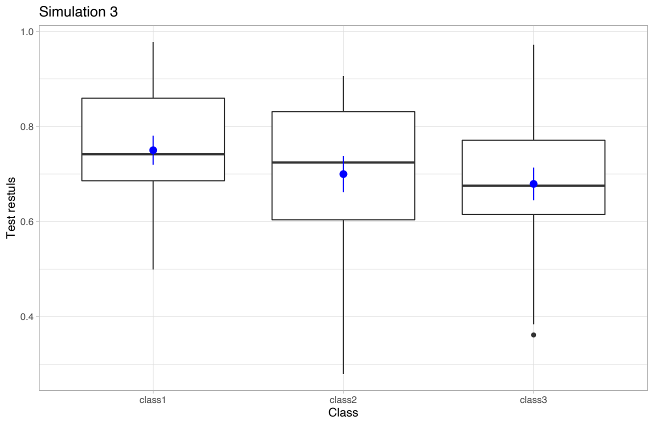
Podemos extrair outras 3 amostras de 20 aluno/as várias vezes, e cada vez os valores de $\overline{X}$ e S mudam (mas μ e σ permanecem os mesmos)

```
   class  Test.mean Test.SD
1 class1      0.683   0.157
2 class2      0.786   0.118
3 class3      0.679   0.158
```



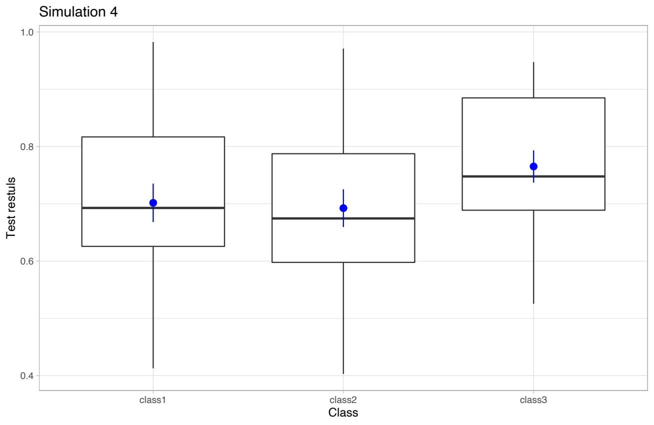
Simulando amostras

```
   class  Test.mean Test.SD
1 class1      0.750   0.137
2 class2      0.700   0.170
3 class3      0.679   0.153
```



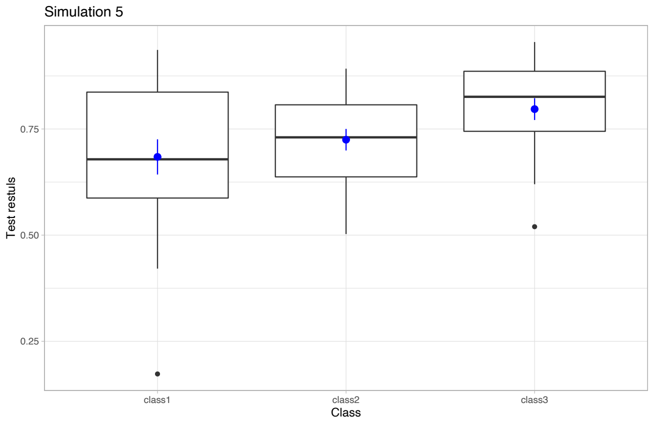
Simulando amostras

```
   class  Test.mean Test.SD
1 class1      0.702   0.150
2 class2      0.692   0.147
3 class3      0.765   0.126
```



Simulando amostras

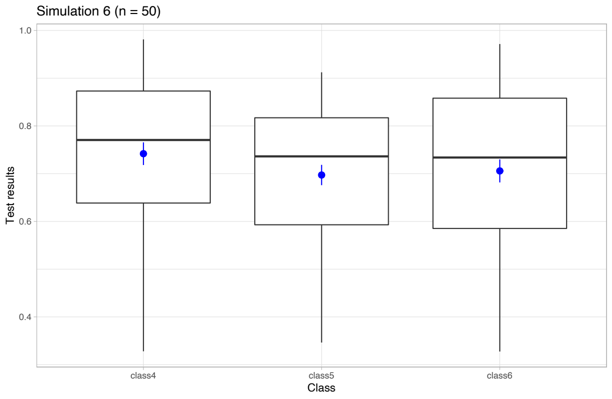
```
  class  Test.mean Test.SD
1 class1     0.684   0.185
2 class2     0.725   0.114
3 class3     0.797   0.115
```



Simulando amostras

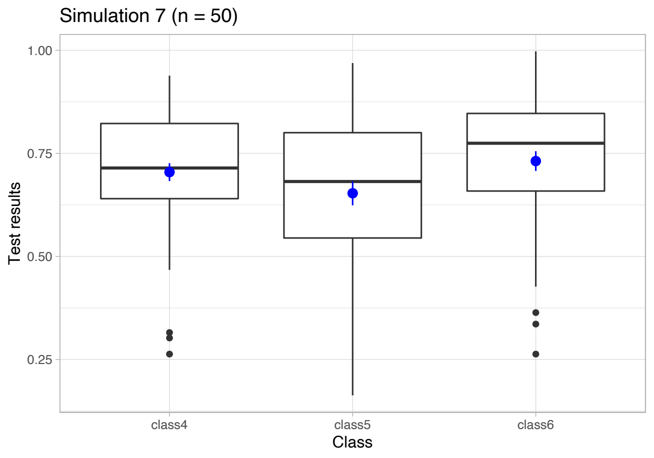
E se extrairmos amostras com 50 aluno/as?

```
    class  Test.mean Test.SD
  1 class4     0.742   0.168
  2 class5     0.697   0.151
  3 class6     0.706   0.170
```



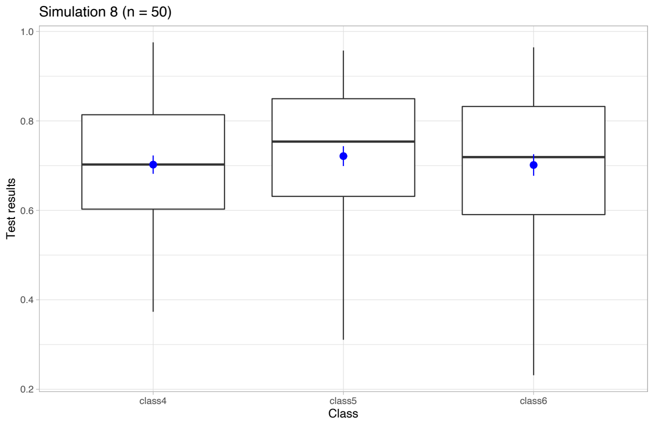
Simulando amostras

```
   class  Test.mean Test.SD
1 class4      0.705   0.154
2 class5      0.653   0.207
3 class6      0.731   0.169
```



Simulando amostras

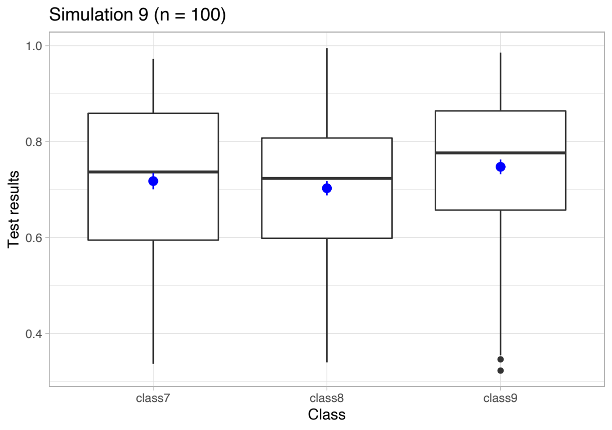
```
   class  Test.mean Test.SD
1 class4     0.702   0.145
2 class5     0.721   0.157
3 class6     0.701   0.170
```



Simulando amostras

E com 100 aluno/as?

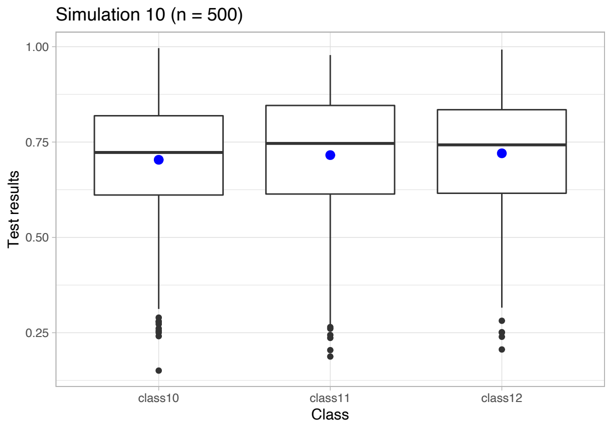
```
   class  Test.mean Test.SD
1 class7      0.718   0.171
2 class8      0.703   0.149
3 class9      0.748   0.154
```



Simulando amostras

E com 500 aluno/as?

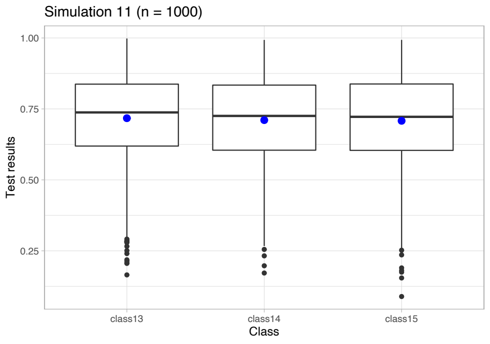
```
  class   Test.mean Test.SD
1 class10     0.703   0.154
2 class11     0.716   0.166
3 class12     0.720   0.152
```



Simulando amostras

E com 1.000 aluno/as?

```
   class  Test.mean Test.SD
1 class13     0.717   0.159
2 class14     0.711   0.159
3 class15     0.708   0.158
```



Lembrando: $\mu = 71,5 \ \sigma = 16$

Conclusão?

- Quanto maior o n, melhor a estimativa dos parâmetros da população,
- mas chega um momento em que não passa mais a fazer diferença.

$\rightarrow$ Como chegar a um n ideal?

- Teste de *poder* estatístico*

cf. Central Limit Theorem & Law of Large Numbers

Valor de p

Noções básicas de probabilidade

- Imagine que vamos fazer uma aposta sobre quem ganha um jogo de cara ou coroa
- Cara eu ganho, coroa você ganha
- Qual é a probabilidade de uma moeda justa cair cara?

$\rightarrow$ 50% ($p = 0.5$)

- Qual é a probabilidade de caírem 3 caras em 3 jogadas se a moeda for justa?
- Opções:
- Ca-Ca-Ca, Ca-Ca-CO, Ca-CO-Ca, Ca-CO-CO
  CO-Ca-Ca, CO-Ca-CO, CO-CO-Ca, CO-CO-CO
- 8 opções $\rightarrow 1/8 = 0.125 \rightarrow$ probabilidade de 12,5% de caírem 3 caras em 3 jogadas

Noções básicas de probabilidade

A probabilidade de caírem 3 caras (ou 3 coroas) em 3 jogadas é de 12,5%.

Sequência	n **de caras**	**probabilidade**
Ca-Ca-Ca	3	0.125
Ca-Ca-CO	2	0.125
Ca-CO-Ca	2	0.125
Ca-CO-CO	1	0.125
CO-Ca-Ca	2	0.125
CO-Ca-CO	1	0.125
CO-CO-Ca	1	0.125
CO-CO-CO	0	0.125

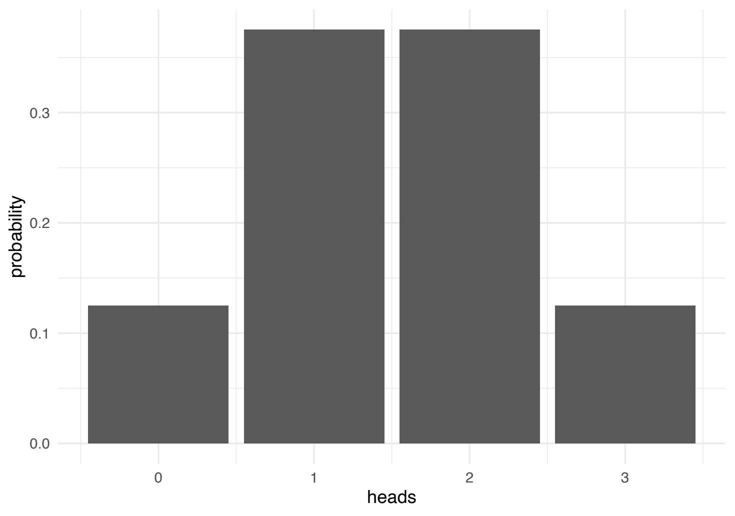
E qual é a probabilidade de caírem 2 caras em 3 jogadas?

Noções básicas de probabilidade

Sequência	n de caras	probabilidade
Ca-Ca-Ca	3	0.125
Ca-Ca-CO	2	0.125
Ca-CO-Ca	2	0.125
Ca-CO-CO	1	0.125
CO-Ca-Ca	2	0.125
CO-Ca-CO	1	0.125
CO-CO-Ca	1	0.125
CO-CO-CO	0	0.125

E qual é a probabilidade de caírem 2 caras em 3 jogadas?
$0.125 + 0.125 + 0.125 = 0.375 \rightarrow 37.5\%$

Ou seja:



Noções básicas de probabilidade

- E se fizermos 6 jogadas, há quantas possibilidades de sequência?

  Ca-Ca-Ca-Ca-Ca-Ca, Ca-CO-Ca-Ca-Ca-Ca, Ca-Ca-CO-Ca-Ca-Ca, Ca-Ca-Ca-CO-Ca-Ca

  ...

  CO-CO-CO-CO-CO-CO
- $2^6 = 64 \rightarrow$ há 64 sequências possíveis
- Em 100 jogadas há $2^{100} = 1.267651e + 30$ $(= 1267651000000000000000000000000)$ sequências possíveis
- Impossível de se calcular as probabilidades a mão!

Noções básicas de probabilidade

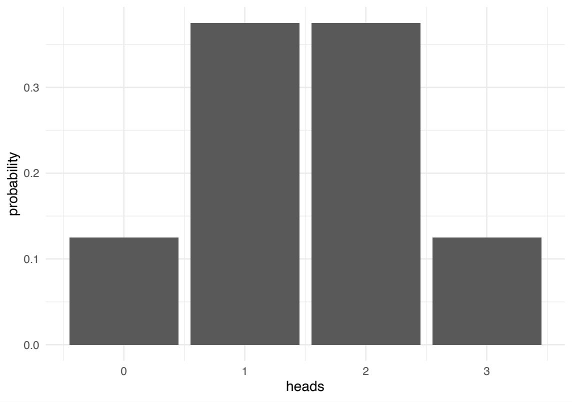
→ R to the rescue!

Podemos utilizar uma distribuição binomial para esses cálculos:

```
# Qual é a probabilidade de caírem 3 caras em 3 jogadas
# onde a probabilidade de cair cara é 0.5?
dbinom(3, 3, 0.5)

[1] 0.125

# Qual é a probabilidade de caírem 0, 1, 2 e 3 caras em 3 jogadas
# onde a probabilidade de cair cara é 0.5?
dbinom(0:3, 3, 0.5)

[1] 0.125 0.375 0.375 0.125

#Plotar as probabilidade
barplot(dbinom(0:3, 3, 0.5)
```

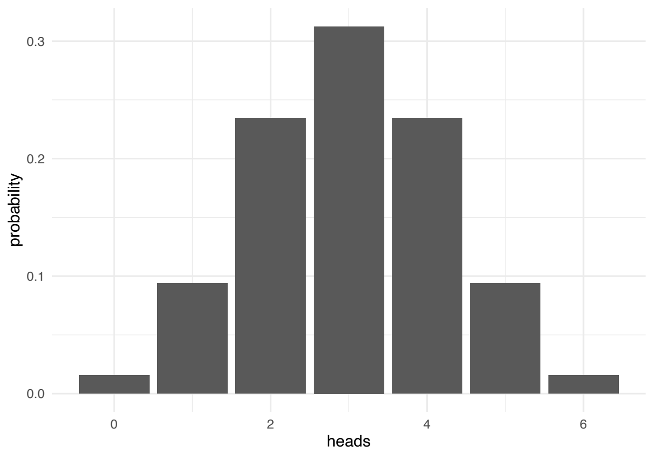
Noções básicas de probabilidade

```
# Qual é a probabilidade de caírem 0, 1, 2 e 3 caras em 3 jogadas
# onde a probabilidade de cair cara é 0.5?
dbinom(0:3, 3, 0.5)

[1] 0.125 0.375 0.375 0.125

#Plotar as probabilidade
barplot(dbinom(0:3, 3, 0.5)
```



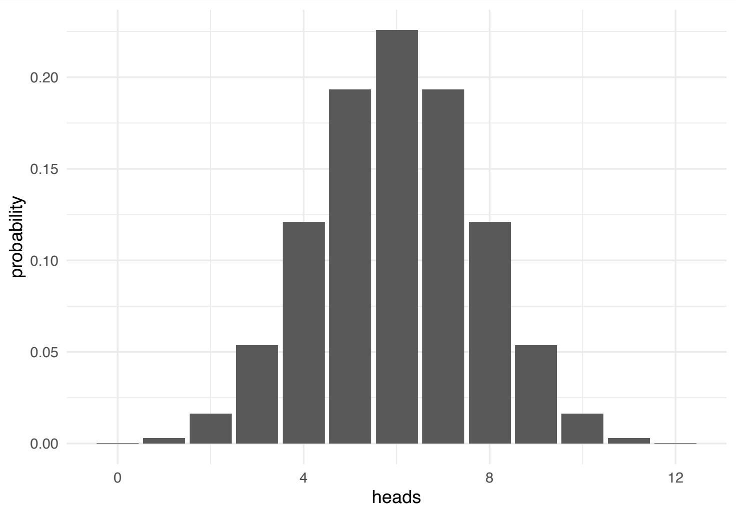
Noções básicas de probabilidade

```
# Probabilidade de caírem 0, 1, 2, ..., 6 caras em 6 jogadas?
dbinom(0:6, 6, 0.5)

[1] 0.015625 0.093750 0.234375 0.312500 0.234375 0.093750 0.015625

barplot(dbinom(0:6, 6, 0.5)
```



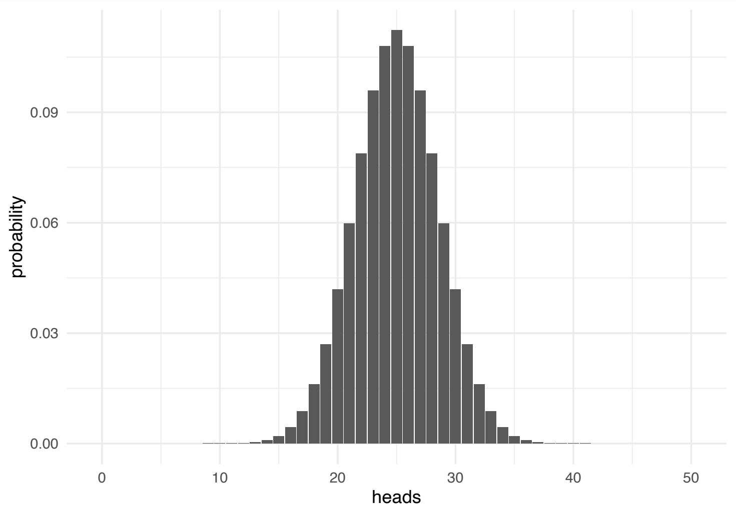
Noções básicas de probabilidade

12 jogadas:



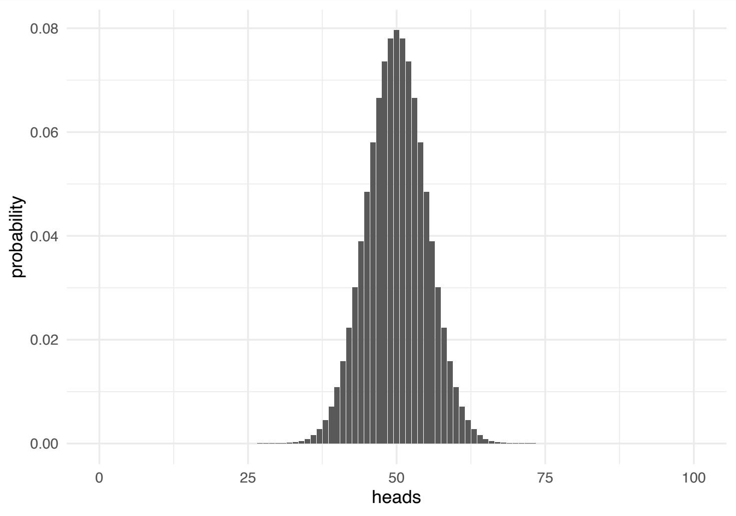
Noções básicas de probabilidade

50 jogadas:



Noções básicas de probabilidade

100 jogadas:



Noções básicas de probabilidade

→ A soma de todas as probabilidade será sempre 1 (100%)

```
sum(dbinom(0:3, 3, 0.5))
[1] 1

sum(dbinom(0:50, 50, 0.5))
[1] 1

sum(dbinom(0:100, 100, 0.5))
[1] 1
```

→ Na nossa aposta de cara eu ganho coroa você ganha que eu propus com a minha moeda, a partir de quantas aparições de cara em 100 jogadas vocês desconfiaria de que eu estou roubando com uma moeda adulterada? *(escreva esse número em algum lugar para comparar depois)*

Noções básicas de probabilidade

→ Com 90 ou mais caras a maioria das pessoas desconfiaria

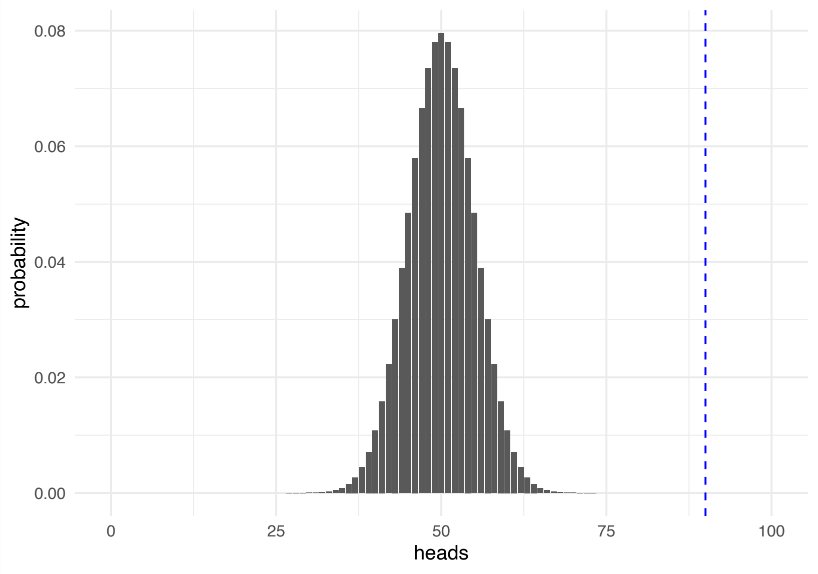
→ Qual é a probabilidade de caírem 90 ou mais caras em 100 jogadas com uma moeda justa?

é a soma da probabilidade de caírem 90 caras com a probabilidade de caírem 91 caras, com a probabilidade de caírem 92 caras, ..., com a probabilidade de caírem 100 caras:

```
sum(dbinom(90:100, 100, 0.5))
[1] 1.531645e-17
```

- A probabilidade (valor de *p*) de caírem 90 ou mais caras em 100 jogadas com uma moeda justa é de 0,000000000000001531645%
- Suficientemente baixa para qualquer um desconfiar de que a moeda não é justa.

Noções básicas de probabilidade



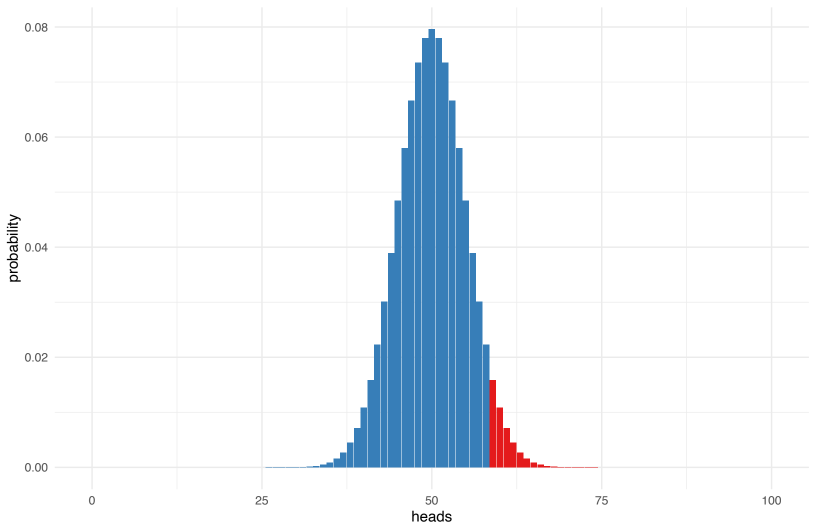
Noções básicas de probabilidade

- Neste caso:
- H_1: A moeda é adulterada (o professor está roubando)
- H_0: A moeda é justa (o professor não está roubando)

$\rightarrow$ Testamos (falseamos) a H_0

- Observamos a quantidade de caras. Se a probabilidade desses dados (a quantidade de caras) for muito baixa sob a H_0, rejeitamos a H_0 (e aceitamos a H_1)
- Mas quão baixa?
- Tradição (arbitrária) é de 5% ($\alpha = 0.05$)
- Então a partir de quantas caras a soma das probabilidades permanece menor que 5%?

Noções básicas de probabilidade



Noções básicas de probabilidade

```
qbinom(0.05, 100, 0.5, lower.tail = F)
[1] 58

sum(dbinom(58:100, 100, 0.5))
[1] 0.06660531

sum(dbinom(59:100, 100, 0.5))
[1] 0.04431304
```

- A probabilidade de caírem 58 ou mais caras é de 6,7% ($p = 0.0666$)
- A probabilidade de caírem 59 ou mais caras é de 4,4% ($p = 0.0443$)

→ Qual foi a quantidade de caras que você anotou como sendo o valor a partir do qual você começaria a desconfiar de mim?

→ Você foi mais ou menos rígida/o do que a tradição de $\alpha = 0.05$ ($p < 0.05$)?

O valor de p

O valor de *p* é o valor da probabilidade

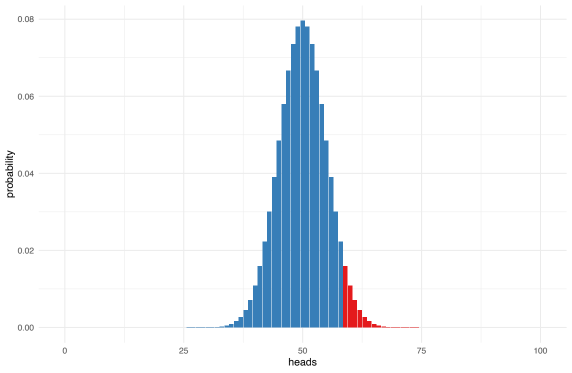
→ **É a probabilidade de dados tão ou mais extremos do que o observado caso a H_0 seja verdadeira**

- No nosso exemplo, é a probabilidade de n ou mais caras caso a moeda seja justa
- Se a essa probabilidade for muito baixa (tradicionalmente < 0.05), desconfiamos de que a H_0 seja verdadeira e a rejeitamos, <u>inferindo</u> (não provando) que a H_1 é verdadeira

Teste unicaudal e bicaudal

Teste unicaudal e bicaudal

Em todos esses exemplos, eu propus a aposta e eu trouxe a moeda, por isso desconfiamos apenas de mim e olhamos apenas para um extremo (uma cauda) do gráfico:

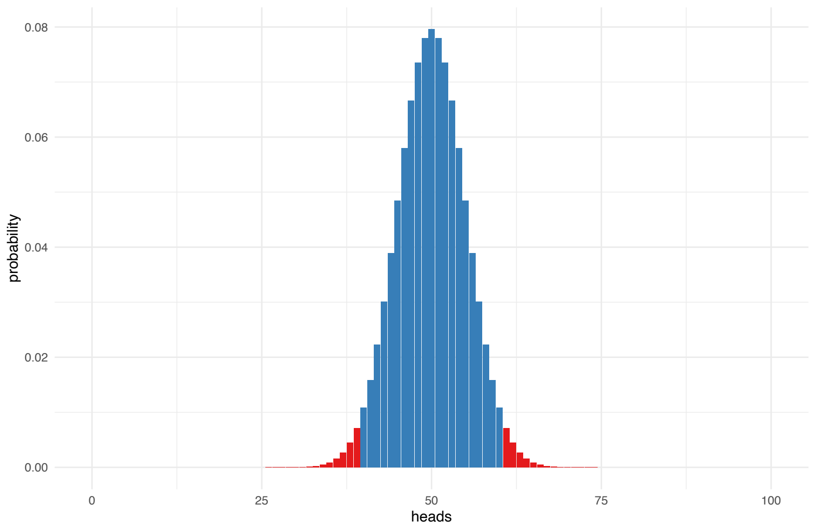


→ Trata-se de um teste unicaudal

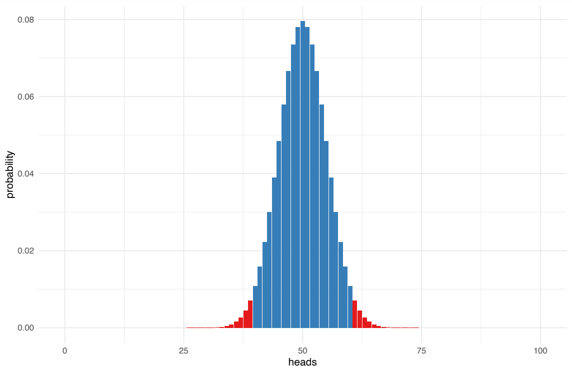
Teste unicaudal e bicaudal

- E se fosse um juiz observando dois apostadores que vão começar a jogar agora, e o juiz não sabe quem trouxe a moeda?
- A princípio, antes de começarem o jogo, ele deve desconfiar dos dois:
- H_1: A moeda foi adulterada (quem está jogando por caras ou quem está jogando por coroas está roubando)
- H_0: A moeda não foi adulterada (é uma moeda justa e ninguém está roubando)
- Ou seja, se caírem caras demais ou caras de menos (i.e., coroas demais), o juiz deve inferir que alguém estava roubando
- Neste caso, o $\alpha = 0.05$ deve ser dividido entre as duas caudas da distribuição (teste bicaudal)

Teste unicaudal e bicaudal



Teste unicaudal e bicaudal



→ Até quantas caras é muito pouco a ponto de se desconfiar do jogador jogando por coroas e a partir de quantas caras é demais a ponto de se desconfiar do jogador jogando por caras?

Teste unicaudal e bicaudal

→ Até quantas caras é muito pouco a ponto de se desconfiar do jogador jogando por coroas e a partir de quantas caras é demais a ponto de se desconfiar do jogador jogando por caras?

```
qbinom(0.05/2, 100, 0.5, lower.tail=FALSE)
[1] 60

sum(dbinom(c(0:40, 60:100), 100, 0.5))
[1] 0.05688793

sum(dbinom(c(0:39, 61:100), 100, 0.5))
[1] 0.0352002
```

→ A probabilidade de caírem 40 ou menos caras mais a probabilidade de caírem 60 ou mais caras é de 5,7% ($p = 0.0569$)

→ A probabilidade de caírem 39 ou menos caras mais a probabilidade de caírem 61 ou mais caras é de 3,5% ($p = 0.0352$)

Teste unicaudal e bicaudal

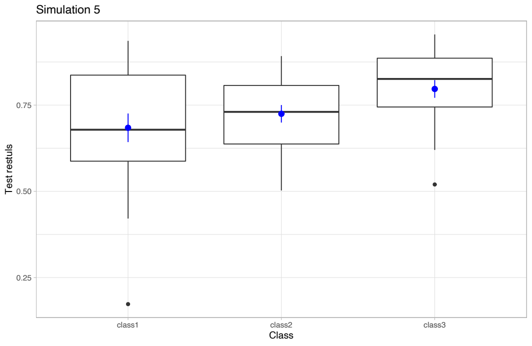
- Em um teste unicaudal (de uma hipótese direcional) com valores menos extremos (59 caras) rejeitamos a H_0

- Em um teste bicaudal (de uma hipótese não direcional) precisamos de valores mais extremos (61 caras) para rejeitar a H_0

Erro de Tipo I e de Tipo II

Amostras simuladas

Voltando às amostras que simulamos, vamos examinar a última simulação que fizemos com 20 aluno/as por turma:



Class	$\overline{X}$	S
1	0.684	0.185
2	0.725	0.114
3	0.797	0.115

Amostras simuladas

- Sabemos que as turmas não devem ser (significativamente) diferentes, pois as amostras vieram da mesma população (que criamos), e conhecemos os verdadeiros parâmetros da população: $\mu = 71,5\ \sigma = 16$
- Mas normalmente temos apenas os dados da amostra e desconhecemos os verdadeiros parâmetros da população
- E se, coincidentemente, essas três turmas fossem

  1 = grupo controle

  2 = grupo experimental 1

  3 = grupo experimental 2

  e quiséssemos inferir o efeito dos tratamentos?

$\rightarrow$ H_1: Há diferença entre as turmas

$\rightarrow$ H_0: Não há diferença entre as turmas

Amostras simuladas

```
summary(aov(data = sample.data, test ~ class))

> summary(aov(data = sample.data, test ~ class))
            Df Sum Sq Mean Sq F value Pr(>F)
class        2  0.130 0.06502   3.227  0.047 *
Residuals   57  1.149 0.02015

TukeyHSD(aov(data = sample.data, test ~ class))

$class
                    diff          lwr       upr     p adj
class2-class1 0.04055251 -0.067467314 0.1485723 0.6403871
class3-class1 0.11258046  0.004560633 0.2206003 0.0392596
class3-class2 0.07202795 -0.035991879 0.1800478 0.2520058
```

- $p < 0.05$: rejeitamos a H_0 e concluímos que houve diferença "estatisticamente significativa" entre as turmas 1 e 3!
- Mas sabemos que não deveria haver
- Isto é um Erro de Tipo I

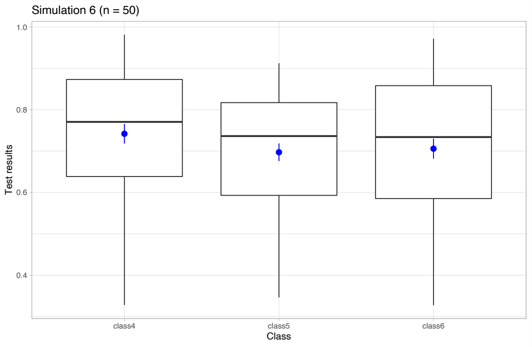
Erro de Tipo I e de Tipo II

Hipótese nula é:	Verdadeira	Falsa
Rejeitada	Erro de tipo I	Decisão correta
Não rejeitada	Decisão correta	Erro de tipo II

- **Erro de Tipo I**: Alarme falso, falso positivo. Concluir como resultado "estatisticamente significativo" quando o efeito apareceu ao acaso
  - probabilidade de cometer erro de Tipo I $= \alpha$ (5%)
- **Erro de Tipo II**: Falso negativo. Não conseguir detectar um efeito quando de fato existe um
  - pode acontecer por baixo *poder estatístico*
  - é possível diminuir as chances desse erro aumentando o tamanho da amostra

Aumentando n para evitar erro de Tipo I

Vamos examinar a primeira simulação que fizemos com 50 aluno/as por turma:



Class	$\overline{X}$	S
4	0.742	0.168
5	0.697	0.151
6	0.706	0.170

Aumentando n para evitar erro de Tipo I

```
summary(aov(data = sample.data2, test ~ class))

> summary(aov(data = sample.data2, test ~ class))
             Df Sum Sq Mean Sq F value Pr(>F)
class         2  0.056 0.02808   1.055  0.351
Residuals   147  3.911 0.02661
```

→ Desta vez não rejeitamos a hipótese nula, e, como sabemos que a H_0 é verdadeira (porque criamos a população e conhecemos seus parâmetros), tomamos a decisão correta.

Erro de Tipo I e de Tipo II

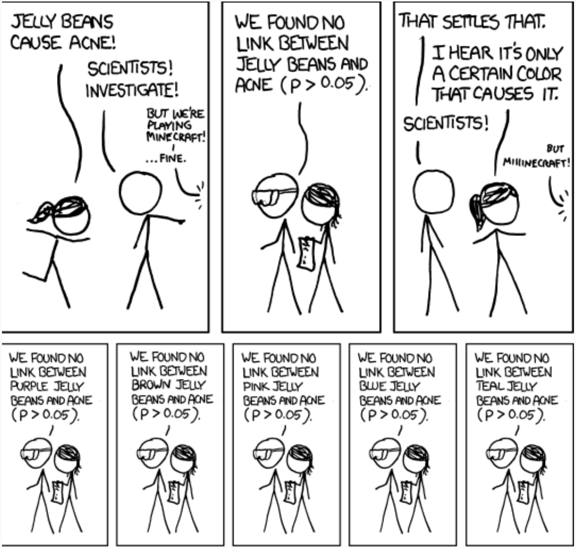
- Diminuir α diminui as chances de erro de Tipo I, mas aumenta as chances de erro de Tipo II
- Aumentar o poder de um teste estatístico diminui as chances de erro de Tipo II, mas aumenta as chances de erro de Tipo I
- Repetir um experimento diversas vezes aumenta a chance de erro de Tipo I. Ao repetirmos um experimento 100 vezes, cometeremos um erro de Tipo I em média 5 vezes com $\alpha = 0.05$

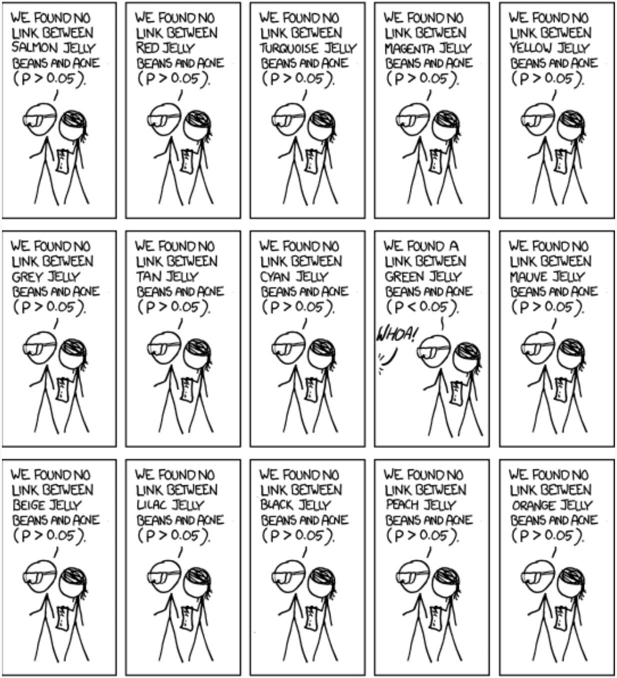
  (como saber se o nosso único experimento não está entre esses 5?)

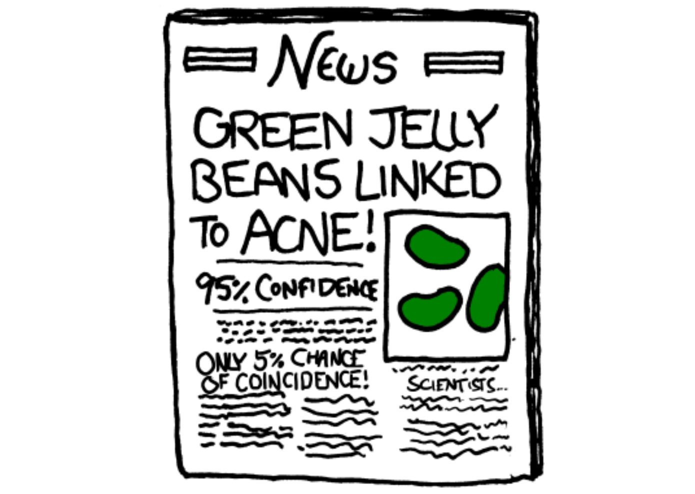
- Algum desses dois tipos de erro é mais grave?

  É pior dizer que há efeito quando não há ou não conseguir identificar um efeito quando há?

https://xkcd.com/882/




WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Cuidados sobre o valor de p

Equívocos comuns sobre o valor de p

O valor de *p*:

- não é a probabilidade da H_0 ser verdadeira (é a probabilidade dos dados diante da H_0)
- não prova que a H_1 seja verdadeira, apenas indica a decisão de rejeitar a H_0 (e aceitamos, por inferência, a H_1)
- não indica a magnitude ou importância de um efeito – um *p* muito baixo não indica um efeito muito alto

$\rightarrow$ Bônus: valor de *p* "marginalmente significativo" (e.g., $p = 0.06$) não indica tendência de efeito/de diferença

Críticas ao valor de p

(e.g., Wagenmakers 2007, Nuzzo 2014, Halsey 2015)

- decisão categórica que valor de *p* impõe
- arbitrariedade do 0,05 como valor limite para decisão
- possibilidade de se manipular os dados a fim de se alcançar um valor de *p* abaixo de 0,05 (*p-hacking*)
- existência de estudos com valor de *p* abaixo de 0,05 mas com baixo poder estatístico e/ou baixo tamanho de efeito
- o valor de *p* apresenta apenas a probabilidade dos dados diante de uma H_0, mas não é capaz de informar sobre a probabilidade da H_1 e do efeito

Alternativas

- Diminuir o foco no valor de p e não depender apenas dele para a inferência. Investigar e reportar
  - tamanho do efeito
  - intervalos de confiança
  - poder estatístico
- Priorizar modelos estatísticos em vez de testes de hipótese
- Utilizar modelos estatísticos que nem mesmo utilizam valores de p (estatística bayesiana)

Perguntas?