MAD-CB

UNIFESP

UNIVERSIDADE FEDERAL DE SÃO PAULO

1933

Inferência – 1 (Parte B)

Distribuição de Amostra

- O que observamos é uma distribuição de amostra
- Nosso trabalho é avaliar a congruência dela com uma distribuição teórica
- Valores observados variam de amostra em amostra
- Esta variabilidade se chama: variância amostral
- Podemos fazer várias amostras e criar uma distribuição das médias ($\bar{x}$)
- Distribuição das amostras terá uma média e variância também

Valor Esperado e Variância da Distribuição de Amostras

- Esses existem por causa da Teorema de Limite Central

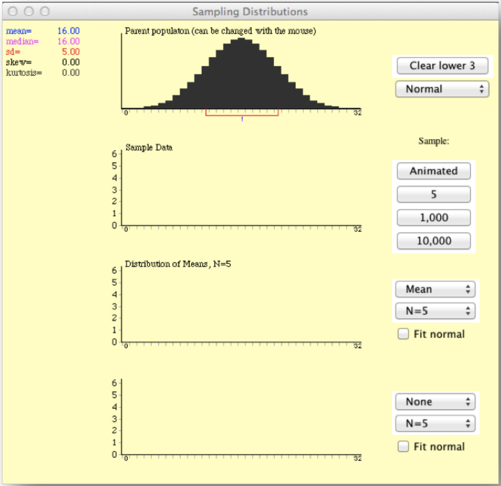
$$E(\bar{X}) = \mu$$

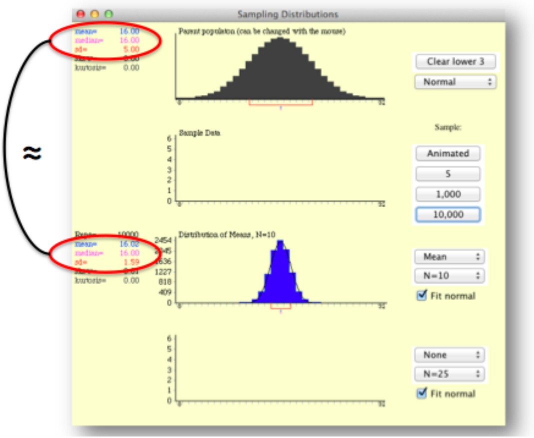
$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

$$DP(\bar{X}) = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

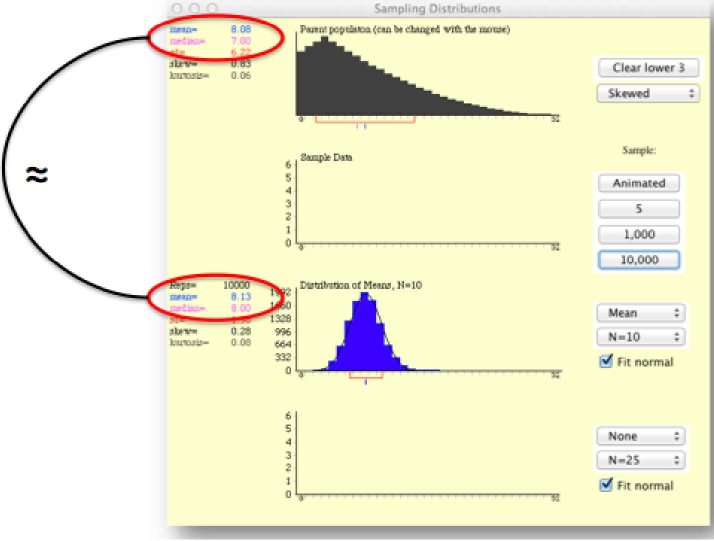
Comparar Estatísticas das Amostras a População

- Rice University – Applet das Distribuições Amostrais -Site: `http://onlinestatbook.com/stat_sim/sampling_dist/index.html`

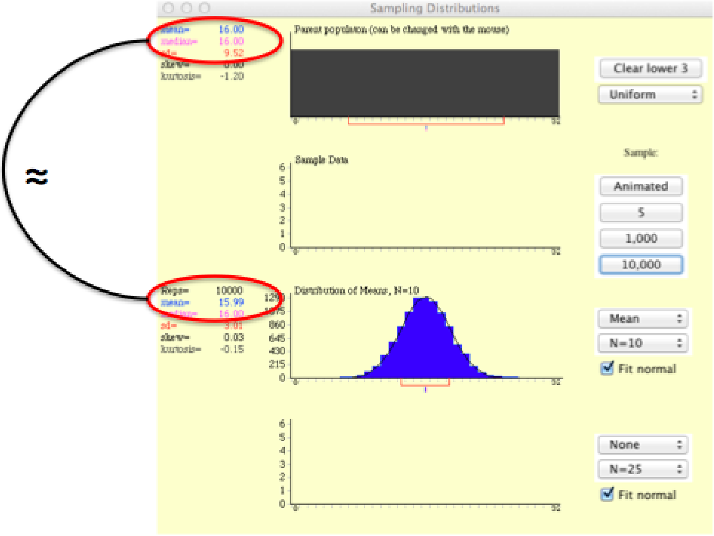




$E(\bar{X}) = \mu;\ 16,02 \approx 16,00$
$DP(\bar{X}) = \frac{\sigma}{\sqrt{n}};\ \frac{5.00}{\sqrt{10}} = 1.58 \approx 1.59$


≈
Sampling Distributions
mean= 8.08
median= 7.00
kurtosis= 0.06
Parent population (can be changed with the mouse)
Clear lower 3
Skewed
Sample:
Sample Data
Animated
5
1,000
10,000
Reps= 10000
mean= 8.13
median= 8.00
skew= 0.28
kurtosis= 0.08
Distribution of Means, N=10
Mean
N=10
Fit normal
None
N=25
Fit normal

Distribuição Uniforme



Resumo - Distribuição Amostral – Proporções

- Teorema de Limite Central (CLT)
- Estudamos amostras e comparar nossa amostra a todas as amostras possíveis
- Distribuição Amostral de proporção binomial

$$\hat{p} \approx N(p, \frac{p(1-p)}{n})$$

```
- Distribuição Amostral da Média
```

$$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

- N.B. $N(\mu, \sigma^2)$ quer dizer distribuição normal com média de μ e variância de σ^2

Vamos Imaginar que Temos uma Garrafa Cheio de Contas

- 2 Cores – Vermelho e Azul
- Não sabemos a proporção de cada cor
- Podemos fazer um experimento
    - Tirar 25 contas da garrafa e contar as cores para estimar a proporção verdadeira
    - Pode repetir isso múltiplas vezes (**muitas**!!) para estimar a proporção na garrafa
    - Usar a função `sample` em R
- *Simulação Monte Carlo*
    - Simular com o computador um evento e repetir muitas vezes
    - Estimação do valor de população
    - Aproveita da Lei de Grandes Números

- Vamos criar as contas com `rep`
  - Vai criar um vetor com todos as contas na garrafa
  - Não vou mostrar aqui
- Vamos selecionar 1 conta da garrafa

```
## [1] "azul"
```

- De novo

```
## [1] "vermelho"
```

Repetir Multiplas Vezes – com `replicate`

- Muitas vezes – 10.000

```
trials <- 10000
set.seed(1)
eventos <- replicate(trials, sample(conta, 1))
head(eventos)
```

```
## [1] "vermelho" "vermelho" "azul"     "azul"     "vermelho"
```

Determinar o Resultado da Simulação

- Usar funções `table` e `prop.table`
  - `table` – tabula os resultados
  - `prop.table` – calcula as proporções dos resultados

```
(tab <- table(eventos))
```

```
## eventos
##     azul vermelho
##     4704     5296
```

```
prop.table(tab)
```

```
## eventos
##     azul vermelho
##   0.4704   0.5296
```

Proporções Verdadeiras

- Divulgação das proporções verdadeiras
  - **azul** – 0.474
  - **vermelho** – 0.526

Com e Sem Substituição

- `replicate` funciona *com substituição*
  - Tirar a conta da garrafa e repor depois
- *Sem substituição* quer dizer que não repormos a conta
  - Fica permanentemente perdido para as tabulações futuras

Distribuições de Probabilidade

- Distribuições dos números e das probabilidades são vinculados
  - Ex: Quincunce
- *Função densidade de probabilidade* $f(x) = c$
  - probabilidade que a distribuição assume um valor específico
- *Função de probabilidade cumulativa* $F(x) \leq c$
  - proporção dos valores na distribuição que ficam abaixo ou igual a um valor específico

Aplicar para Proporção das Contas Azuis

- Converter as cores em números (“azul” $= 1$)

```
contnum <- as.numeric(conta == "azul")
```

- Podemos acertar que o função cumulativa para “azul” (1)
  - $F(1) = \frac{474}{1000} = 0.474$
- Para “vermelho” (0)
  - $F(0) = \frac{526}{1000} = 0.526$

Para Variáveis Categoricas – Distribuição Cumulativa Não Intuitiva

- Melhor fazer o que fizemos com os números
- Define probabilidade de todos os estados possíveis da variável $\Pr(\text{vermelho}) = 0.526$ e $\Pr(\text{azul}) = 0.474$.

Inferência - 2

Motivação – Sondagens Políticas

- Quando IBOPE diz que um candidato está em frente do outro por 52% a 48% com uma *margem de erro* de 4%, o que quer dizer essa margem de erro?
- Conceito de margem de erro implica que as variáveis são aleatórias
- Daí pode tratar dos assuntos de:
  - Intervalos de confiança
  - Valor p

Tirando Conclusões das Proporções – Exemplo

- Uma cidade tem exatamente 1.000.000 eleitores
  - 504.000 Republicans
  - 496.000 Democrats
- Pesquisador chega para fazer uma sondagem
  - Questão – Quantas Democrats tem a cidade?
- Não sabe o valor da população (49,6%)
- Quer estimar este valor através amostras

A Sondagem

- Sonda afiliação partidária de uma amostra de 1000 eleitores aleatórios

```r
poll <- sample(cidade, npoll, replace = TRUE)
table(poll)
```

```
## poll
##   D   R
## 498 502
```

- Previsão da sondagem é vitoria para os Republicans
- Mas, esta amostra representa a população?
- A *estimativa* do resultado reflete a realidade?

Variáveis Aleatórias

- Os resultados dos processos aleatórios
- A sondagem selecionou 1% dos eleitores aleatoriamente
- O que acontece se fazemos isso várias vezes (5)
- Vamos contar os Democrats em 5 sondagens
- Resultados de 5 sondagens: 479, 493, 509, 492, 513
  - Em algumas, os Democrats ganham
- Pode ver que resultados variam bastante
  - Variância *aleatória*
- Para entender os resultados, precisa entender **modelos de amostragem**

Modelos de Amostragem

- Qual valor podemos esperar de nossa sondagem original?
  - Probabilidade de ser Democrat (p = 0.496) x tamanho de amostra (npoll = 1000)

$$E(Dem) = 1000p$$

```
evDems <- p * npoll
```

- Valor Esperado ($E(Dem)$) = 496

Erro Padrão

- Mostra tamanho do erro aleatório
- Erro Padrão dos valores

$$\mathrm{SE}(Dem) = \sqrt{1000p(1-p)}$$

```
seDem <- sqrt(npoll * p * (1 - p))
```

- Erro padrão dos valores $= 15.811$
  - Erro fica mais ou menos 496 ± 15.811

Versão Normalizada

- Pode normalizar esses valores controlando para tamanho de amostra
- Valor esperado da proporção na amostra

$$E(Dem/1000) = p$$

- Este implica que $Dem/1000$ mais um erro aleatório igualará à p

Erro Padrão da Proporção

- Dá um tamanho mais exato a correção necessário na amostra

$$\text{SE}(Dem/1000) = \frac{\sqrt{p(1-p)}}{\sqrt{N}}$$

```
sePad <- sqrt(p * (1 - p)/sqrt(npoll))
```

- $\text{SE}(Dem/1000) = 0.089$

Erros e Tamanho de Amostra

$$\text{SE}(Dem/1000) = \frac{\sqrt{p(1-p)}}{\sqrt{N}}$$

- O que acontece se aumentamos o tamanho de amostra (N)?

Estimativas

- $Dem/1000$ é nossa estimativa de p
- Notação $\hat{p} \approx p$
- O valor esperado exato depende do valor de p que não sabemos
- Melhor aproximação para p é $\hat{p}$
- Assim, podemos dizer que

```
p_hat <- mean(poll == "D")
se <- sqrt(p_hat * (1 - p_hat)/1000)
cat("Nossa estimativa da proporção dos Democrats\né", p_hat,
    "mais ou menos", round(se, 5))
```

```
## Nossa estimativa da proporção dos Democrats
## é 0.498 mais ou menos 0.01581
```

Distribuição de Probabilidade para as Variáveis Aleatórias

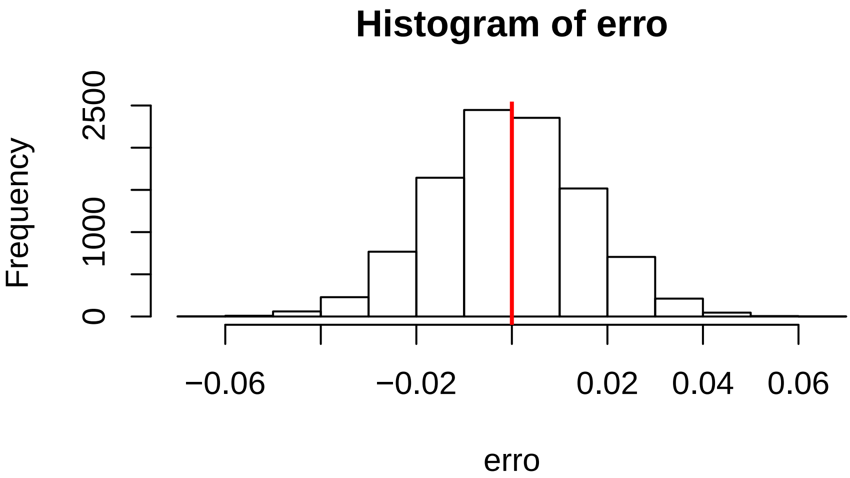
- O "mais ou menos" não é muito útil
- Podemos calcular a probabilidade que $\hat{p}$ fica dentro de 1% do verdadeiro p?
- Vamos começar com uma simulação de nossas eleitores
- Medir a distribuição dos erros $\hat{p} - p$

```
trials = 10^4
erro <- replicate(trials, {
  X <- sample(cidade, npoll, replace = TRUE)
  mean(X == "D") - p
})
```

```r
mean(abs(erro) > 0.01581) ## erros maiores que o SE
```

```
## [1] 0.3246
```

```r
hist(erro)
abline(v = 0.0, col = "red", lwd = 2)
```

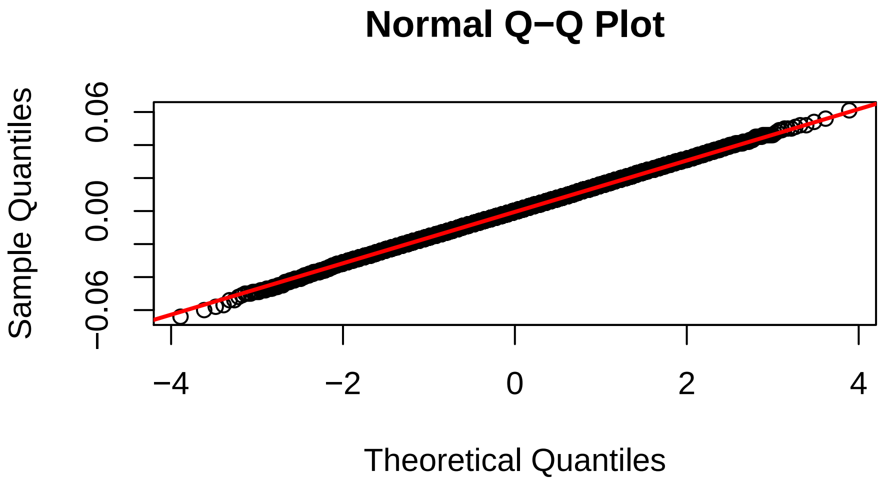


Implicações da Histograma

- Esta é a distribuição de probabilidade de nossa sondagem
- Distribuição da $\hat{p}$ parece perto a normal
- Centro da distribuição em 0
    - Confirma que valor esperado de $\hat{p}$ é p

Confirmação de Aproximação à Normal

```r
qqnorm(erro)
qqline(erro, col = "red", lwd = 2)
```



Comparação dos Dados com a Distribuição Normal

- Comparar % dos erros maior que o SE

```
cat("Proporção verdadeira: ", mean(abs(erro) > 0.01581))
```

```
## Proporção verdadeira:  0.3246
```

- à proporção prevista pela distribuição normal

```
cat("Proporção teorica: ", pnorm(-1) + (1 - pnorm(1)))
```

```
## Proporção teorica:  0.3173105
```

Conclusão

- Podemos dizer em conclusão:

  Com só uma sondagem, podemos dizer que nossa estimativa da proporção de Democrats é $\hat{p}$ e há uma chance de 32% que nosso erro fica maior de que 1.581%

Intervalos de Confiança

- Varição aleatória faz a sondagem não acerta o valor correto 32% das vezes
  - Para uma empresa de sondagens, não muito bom
- Se falamos de um intervalo que acerta 95% das vezes, estamos bem pensados no mercado
- Podemos construir um intervalo $[A, B]$ em que:

$$\Pr(A \leq p \text{ and } B \geq p) \geq 0.95$$

Nota de Rodapé – Porque 0.95

- **Costume Tradição**
- Não tem mágica teórica

Variável Aleatória Z

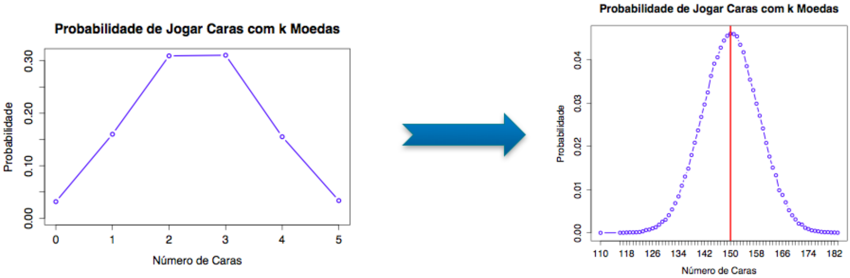
- Como escolhemos A e B para fazer este intervalo tão pequeno quanto possível?
- Sabemos que $\hat{p}$ segue uma distribuição normal (por causa da CLT) com valor esperado de p e erro padrão de $\sqrt{\hat{p}(1-\hat{p})}/\sqrt{N}$
- Esse implica a variável aleatória seguinte (Z):

$$Z = \sqrt{N}\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}}$$

- Z é aproximadamente normal com
  - Valor esperado de 0
  - Desvio padrão de 1

Teorema de Limite Central (CLT) - Repeteco

- Se repetimos um experimento muitas vezes, a probabilidade do resultado médio irá convergir a uma distribuição normal (curva de sino)
- Permite que usamos a distribuição normal como base da maioria de nossos testes estatísticas (paramétricas)



Para CLT Funcionar – Premissas Requisitadas

- Amostras são aleatórias
- Observações são independentes
  - Nenhum tem relação com nenhum outra
- Dados são corretos
  - Neste caso, as pessoas falam a verdade; não mentem
  - Grande problema com sondagens politicas
  - Também auto-descrições das sintomas por pacientes

O Sábio Dr. House

- Todos Mentem

IC – Exemplo

- Mais uma jogada com moedas
  - Jogar 1 moeda 1.000 vezes
  - Usando uma simulação Monte Carlo
- Queremos descobrir a verdadeira, mas desconhecida probabilidade (p) de jogar CARA
  - Fazer com números: CARA $=1$; COROA $=0$

```
set.seed(1); n <- 1000; k <-  1; prob <- 0.5
tiras <- rbinom(n, k, prob)
(caras <- sum(tiras)) ## número de CARAS
```

```
## [1] 480
```

- Fazemos estimativa de p com a amostra de 1.000 jogadas $\hat{p} = 480/1000 = 48\%$
- Com qual grau de confiança podemos dizer que o valor da população p é realmente perto a nosso estimativa da proporção das CARAS?

Equações para Intervalo de Confiança das Proporções

- Sabemos a média (valor esperado) e variância da proporção estimada - $\hat{p}$

$$E(\hat{p}) = \frac{480}{1000} = 0.516$$

$$Var(\hat{p}) = \frac{p(1-p)}{n}$$

- E, por causa da CLT, sabemos que

$$\hat{p} \approx N(p, \frac{p(1-p)}{n})$$

Equações para Intervalo de Confiança das Proporções – 2

- Podemos converter os valores em uma contagem Z
  - Normalizar os valores em termos da média e desvio padrão

$$z_i = \frac{x_i - \bar{x}}{s}$$

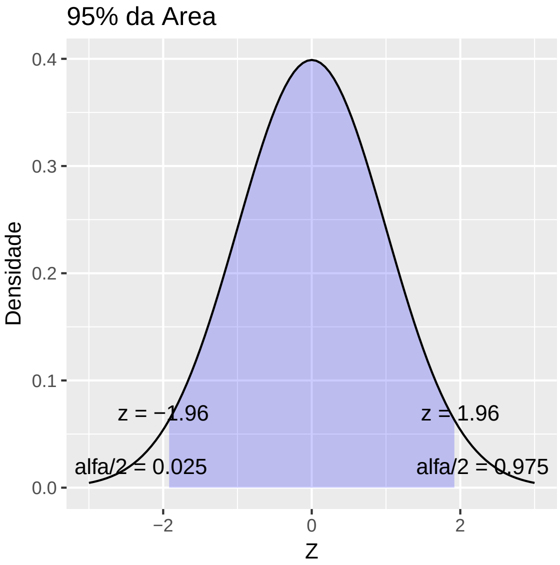
- Contagem Z vem da distribuição normal padronizada, que tem $\mu = 0$ e $\sigma = 1$

$$z = N(0, 1)$$

- Podemos substituir nossos valores nessas equações

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1)$$

Distribuição Normal Padronizada



O Que Significa Isso?

- Para 95% das amostras, z vai ficar entre -1.96 e 1.96
- Valores mais extremos que esses vão ocorrer só 5% das vezes
- Região em que estamos confiantes que nosso valor $\hat{p}$ representa o valor da população verdadeira
  - -1.96 é o limite inferior
  - 1.96 é o limite superior -Temos 95% confiança que o valor verdadeiro desconhecido de p fica dentro deste intervalo
- $\therefore$ "Intervalo de Confiança"
- Probabilidade que nosso $\hat{p}$ cai fora deste intervalo é só 5% ou menos
- 19 de 20 amostras vai ter um p que cairia dentro do intervalo e só 1 vai ter um valor fora

Formula para Intervalo de Confiança

95% Intervalo de confiança de p:

$$\left[\hat{p} - 1,96\sqrt{\frac{p(1-p)}{n}}, \quad \hat{p} + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$$

3 Elementos para Calcular Intervalo

95% Intervalo de confiança de p:

$$\left[\hat{p} - 1,96\sqrt{\frac{p(1-p)}{n}},\ \hat{p} + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$$

Proporção Estimada

```
R> phat <- sum(flips)/1000
```

Valor Crítico $Z_{\alpha/2}$

```
R> z = qnorm(nivel/2, mean=0, sd=1, lower.tail=FALSE)
```

Margem de Erro

```
R> marg.erro = z * sqrt(phat*(1-phat)/1000)
```

Usamos esses 3 elementos no cálculo do intervalo

Calcular Um IC para Proporção

```
phat <- sum(tiras)/1000
nivel <- 0.05
z <-  qnorm(nivel/2, mean = 0, sd = 1, lower.tail = FALSE)
marg.erro  <-  z * sqrt(phat*(1 - phat)/1000)
(ci <- phat + c(-marg.erro, +marg.erro))
```

```
## [1] 0.4490351 0.5109649
```

- Nosso estimativa de $\hat{p}$ (0.48) cai dentro do intervalo. Serve como boa estimativa

Calcular um IC Usando Pacote `binom`

- Facilita cálculos com a distribuição binomial

```
##        method   x    n mean     lower     upper
## 1 asymptotic 480 1000 0.48 0.4490351 0.5109649
```

Testes de Hipoteses das Médias

Exemplo – Temperatura Normal Humana

- Temperatura normal dos seres humanos usualmente dada como 37ºC
- É verdade? Fazemos um teste empírico com 130 ~~cobaias~~ alunos
  - Alunos canadenses neste caso

```
temps <- read_table("TempData.txt", col_names = FALSE)
colnames(temps) <- "tempC"
suppressMessages(library(psych))
psych::describe(temps)
```

```
##       vars   n  mean   sd median trimmed  mad   min   max range skew
## tempC    1 130 36.81 0.41  36.83   36.81 0.41 35.72 38.22   2.5    0
##       kurtosis   se
## tempC     0.65 0.04
```

Resumo das Estatística da Amostra

```r
xbar <- mean(temps$tempC); paste ("Média =", xbar) # média
```

```
## [1] "Média = 36.8051282051282"
```

```r
dp <- sd(temps$tempC); paste ("Desvio Padrão =", dp) # desvio padrão
```

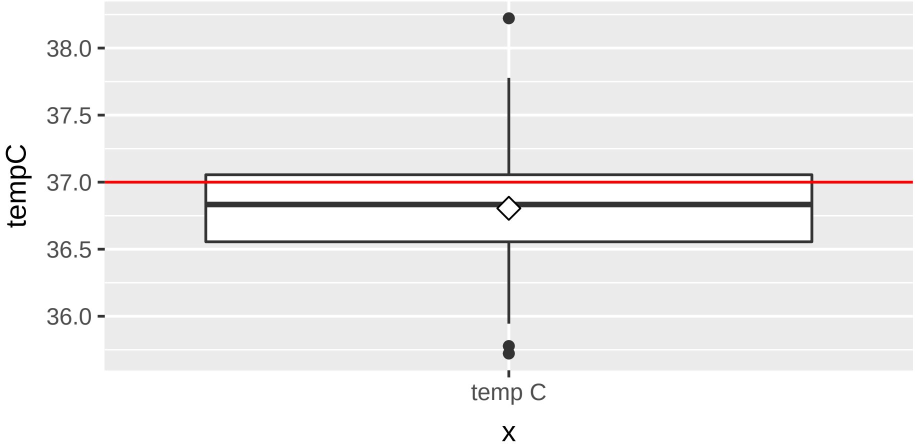
```
## [1] "Desvio Padrão = 0.407323976688302"
```

```r
n <- length(temps$tempC); paste ("n =", n)
```

```
## [1] "n = 130"
```

Boxplot da Amostra

```
boxtemp <-  ggplot(temps, aes(y = tempC, x = "temp C")) + geom_boxplot()
boxtemp <- boxtemp + geom_hline(yintercept = 37, color = "red")
boxtemp <- boxtemp + stat_summary(fun.y="mean", geom="point",
                                  shape=23, size=3, fill="white")
boxtemp
```



Perguntas

- Lembrete dos Números Chaves
  - Média da Amostra: 36.8
  - Normal teórica: 37
- Qual é a probabilidade de obter diferenças de 0.2 graus?
- A diferença entre 36.8 e 37.0 é significativa?

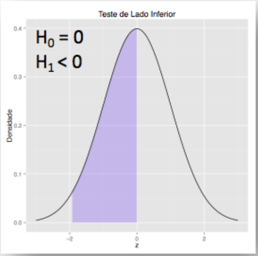
Testes de Hipoteses

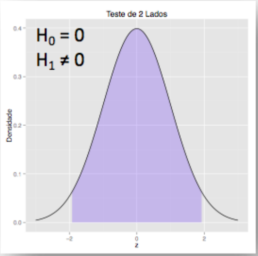
- Testes de contradição
- Não podemos provar diretamente uma hipótese
- Precisamos derrubar uma hipótese que podemos testar
- E ter uma alternativa na mão
- Estamos trabalhando com incerteza e variabilidade natural
- Nós vamos procurar uma resposta testando nossos dados contra o mundo teórico das distribuições

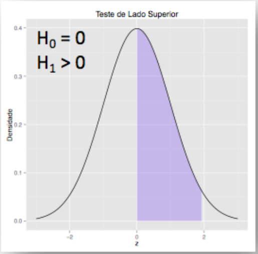
Passos para Formulação e Execução dos Testes

1. Formular uma hipótese (e alternativa) e desenhar um teste
    - Hipótese que vamos testar é a "hipótese nula": H_0
    - Hipótese alternativa é a hipótese de pesquisa: H_1
    - Vamos ver se tivermos suficiente evidência para negar H_0
    - Podemos conduzir o teste de um lado ou de dois lados da distribuição

As Três Condições







2. Colecionar dados e calcular estatística de teste

- Dados calculados baseado na ideia que H_0 é verdade

3. Transformar estatística de teste na escala probabilística

$$0 \leq p \leq 1$$

- Assumindo H_0, quão provável seria a observação de uma estatística de teste deste porte (ou maior) aleatoriamente
- Menor o valor de probabilidade (p), mais forte é a evidência contra H_0
- H_0 ou é verdade ou não é verdade — não assume valores aleatoriamente
- Valor p avisa quão prováveis seriam os dados observados se H_0 for verdade

4. Formar uma conclusão baseada no valor p

- 2 Escolhas
  - Valor p não é pequeno
  - $\therefore$ *Dados consistente com* H_0
  - Valor p é pequeno
  - $\therefore$ *Dados suficiente fora de normal contra* H_0 *em favor de* H_1 *que o resultado é significativa*
- Quão pequeno é *pequeno*
  - Como na CI, $p \leq 0.05$

Força de Evidência

Probabilidade (Valor *p*)	Força de Evidência
$p \leq 0{,}001$	Muito forte
$0{,}001 \leq p \leq 0{,}01$	Forte
$0{,}01 \leq p \leq 0{,}05$	Moderadamente forte
$0{,}05 \leq p \leq 0{,}1$	Fraco
$p \geq 0{,}1$	fugeddabit

Teste de Média

- Usamos mesmo tipo de cálculo para média que para proporção em termos de dados que precisamos: $\bar{x}, s^2$ e n
    - Lembrete de Distribuição Amostral de $\bar{x}$

$$\text{Distribuição Amostral de } \bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$

- Mas, σ^2 desconhecido

Podemos substituir s^2 para σ^2?

- Sabemos s^2 – variança da amostra
- Quase, mas não
- Em vez disso, precisamos usar uma distribuição semelhante à distribuição normal
- Distribuição t (mais formalmente Student's t)
- Historia de "Student" – William Sealy Gossett de Cervejaria Guinness em Dublin

Estatística Teste para Distribuição t

$$t_{n-1} = \frac{\overline{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

DPM

graus de liberdade

Graus de Liberadade

- Calculo de variância da amostra use $(n-1)$ invés de n
  - Função `sd` em R usa $(n-1)$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

- Grau de liberdade (df) representa os $(n-1)$ desvios que podem assumir valores independentemente
- O último valor deve fazer o total $= 0$; $\therefore$ **não tem liberdade**
- Para teste t, os graus de liberdade são $(n-1)$

Student's t – Família de Distribuições

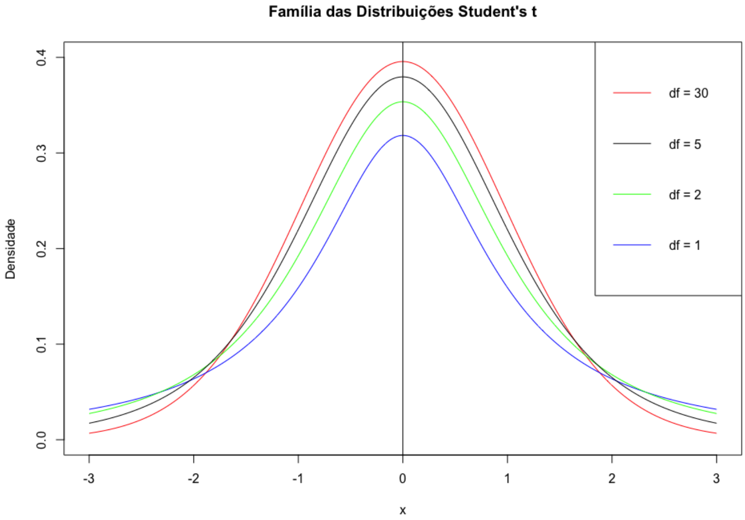
- Cada grau de liberdade define uma curva diferente da distribuição t
- As curvas têm forma semelhante com a curva normal, com caudas mais grossas
- Quando df's aproximam ∞, curva aproxima curva normal
- No exemplo seguinte, pode ver que com uma amostra de 51 e 95% confiança, valores críticos de normal e t ainda são diferentes

```
paste("Valor Crítico -- Normal =", qnorm(0.975))
```

```
## [1] "Valor Crítico -- Normal = 1.95996398454005"
```

```
paste("Valor Crítico -- t Dist =", qt(0.975, 50))
```

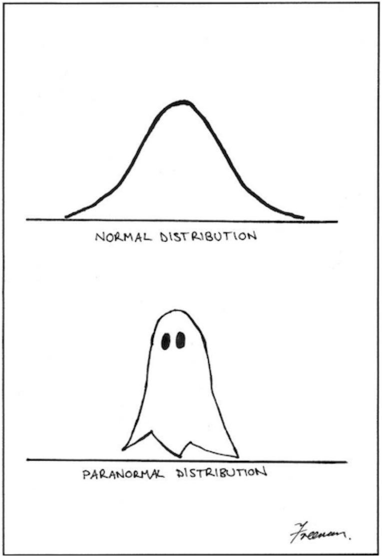
```
## [1] "Valor Crítico -- t Dist = 2.00855911210076"
```


Família das Distribuições Student's t
df = 30
df = 5
df = 2
df = 1
Densidade
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x

Funções das Distribuições em R

- Cada distribuição tem 4 funções que mostram valores associados com ela
  - *d*: densidade (probabilidade de que x vai ter este valor)
  - *p*: área sob a curva da distribuição a esquerda do valor (entre $-\infty$ e o valor)
  - *q*: o valor da distribuição do percentil ou quantil q
  - *r*: números aleatórios usando a distribuição
- R tem muitas distribuições: as mais comunas:
  - Normal (`norm`)
  - Uniforme (`unif`)
  - t (`t`)
  - F (`f`)
  - Binomial (`binom`)
  - Poisson (`pois`)
  - Qui-quadrado (`chisq`)

Existem um Variedade Larga de Outras Distribuições



Funções das Distribuições

- Chamadas às funções tem o formato:
  - [dpqr]<distribuição>
- Exemplos:

```
dnorm(1.96)
```

```
## [1] 0.05844094
```

```
pnorm(1.96)
```

```
## [1] 0.9750021
```

```
qnorm(0.975)
```

```
## [1] 1.959964
```

```
runif(3, 0, 1) ## 3 números aleatórios entre 0 e 1 da dist. Uniforme
```

```
## [1] 0.5308088 0.6848609 0.3832834
```

Teste de Temperatura Normal

- Passo 1: formular as hipóteses
  - $H_0 : \mu = 37$ (hipótese nula que vamos testar)
  - $H_1 : \mu \neq 37$ (hipótese alternativa que é o foco de nossa pesquisa)
  - Teste é de dois lados
  - Usamos um valor crítico de probabilidade de 0.05 (0.025 de cada lado)
- Passo 2: Colecionar Dados e Calcular a Estatística de Teste

```
describe(temps$tempC)
```

```
##    vars   n  mean   sd median trimmed  mad   min   max range skew kurtosis
## X1    1 130 36.81 0.41  36.83   36.81 0.41 35.72 38.22   2.5    0     0.65
##      se
## X1 0.04
```

```
## Estatística de teste
mu <- 37; df <- n - 1
(tstat <- (xbar - mu) / sqrt(dp^2 / n))
```

```
## [1] -5.454823
```

Passos de Exemplo – 2

- Passo 3 – Transformar estatística em probabilidade

```
2 * pt(tstat, df) # para teste de 2 lados
```

```
## [1] 0.0000002410632
```

- Passo 4 – Formar conclusão
  - Valor p é muito pequeno (0.00000024106)
  - Com certeza, abaixo do nível de 0.05 (nosso valor crítico)
  - $\therefore$ vamos rejeitar a H_0 por causa deste valor pequeno
- Interpretação
  - Só rejeitamos a hipótese nula
  - Este não quer dizer que aceitamos a alternativa
  - Só sabemos que a temperatura normal de 37º provavelmente não está totalmente correto

Função de Teste t no R

- R tem uma função que conduz o teste-t sem você precisar calcular a estatística de teste

```r
t.test(temps$tempC, mu = mu, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  temps$tempC
## t = -5.4548, df = 129, p-value = 0.0000002411
## alternative hypothesis: true mean is not equal to 37
## 95 percent confidence interval:
##  36.73445 36.87581
## sample estimates:
## mean of x
##  36.80513
```

Teste t: 2 Amostras

Exemplo — Homicídio Doloso em São Paulo

- Homicídios e tentativas de homicídio subiram muito na percepção pública em São Paulo no último trimestre de 2012
  - depois de uma década de declínio
- Realmente aumentaram?
- 2 amostras medindo os totais mensais dessas categorias em 2011 e 2012
- Baseado nos dados de SSP do Estado de São Paulo
  - Arquivo em formato R: "Crimes.PMSP"
- `d` é a diferença por mês entre 2012 e 2011

Passo 1 – Formular Hipóteses

- $H_0 : d = 0$ (hipótese nula que vamos testar)
- $H_1 : d > 0$ (hipótese alternativa que é o foco de nossa pesquisa)
- Teste unilateral (one-sided: $>$)
- Valor crítico para o teste: $\alpha = 0{,}05$

Passo 2 – Colecionar e se Familiarizar com os Dados

```
load("Crimes.RData")
describe(Crimes.PMSP$TotHD2011)
```

```
##    vars  n   mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 12 179.17 20.25  179.5   179.3 22.24 146 211    65 -0.1    -1.26
##      se
## X1 5.85
```

```
describe(Crimes.PMSP$TotHD2012)
```

```
##    vars  n   mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 12 248.08 50.96    232   243.5 48.18 193 349   156  0.7    -1.03
##       se
## X1 14.71
```

```r
apply(Crimes.PMSP[,8:9], 2, mean)
```

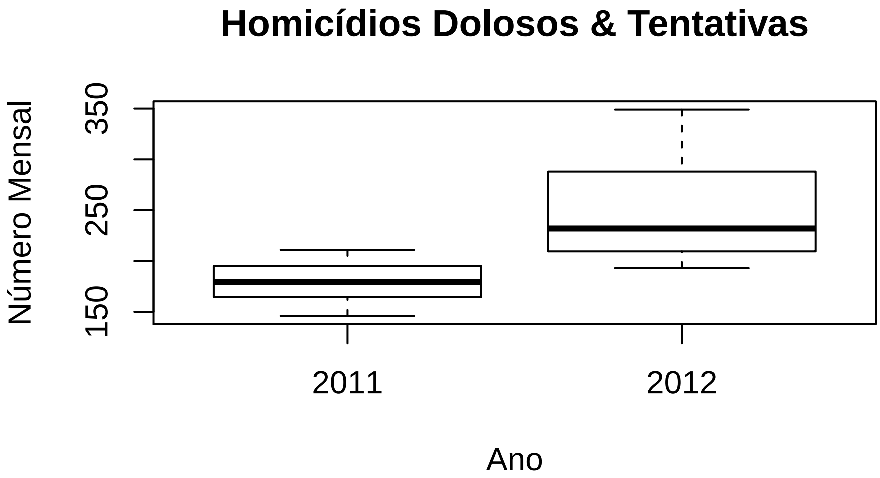
```
## TotHD2011 TotHD2012
##  179.1667  248.0833
```

```r
apply(Crimes.PMSP[,8:9], 2, sd)
```

```
## TotHD2011 TotHD2012
##  20.24771  50.96248
```

Boxplot dos Dados

```
boxplot(Crimes.PMSP[,8:9], horizontal = FALSE, xlab = "Ano",
        ylab = "Número Mensal", main = "Homicídios Dolosos & Tentativas",
        names = c("2011", "2012"))
```



Passo 3 – Teste t

```r
with(Crimes.PMSP, t.test(TotHD2012, TotHD2011, mu = 0, alternative = "greater"))
```

```
##
##  Welch Two Sample t-test
##
## data:  TotHD2012 and TotHD2011
## t = 4.3535, df = 14.388, p-value = 0.0003111
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  41.08784      Inf
## sample estimates:
## mean of x mean of y
##  248.0833  179.1667
```

Passo 4 – Interpretação

- Rejeitar a H_0: a diferença entre 2012 e 2011 foi significativa ao nível de $\alpha = 0.05$
- O que é o valor certo para a população ainda não sabemos, mas sabemos que é maior que 0
- O teste de 2 amostras independentes que fizemos é a versão mais geral dos testes t

Anotações

- `mu = 0`: valor sendo testado é a diferença entre as duas médias (d)
- `alternative = "greater"` : linguagem para um teste unilateral
- `df = 14.388` : porque os tamanhos de amostras não são iguais, calculo dos graus de liberdade precisa contabilizar esta diferença
- `p-value = 0.0003111` : valor abaixo o valor crítico de $\alpha = 0.05$

Exemplo 2: Expectativa da Vida por Região do Mundo

- Comparação dos países das Américas com África Subsaariana
- Formular as Hipóteses
  - $H_0 : d = 0$ (hipótese nula que vamos testar)
  - $H_1 : d \neq 0$ (hipótese alternativa que é o foco de nossa pesquisa)
- Estamos testando a ideia que a expectativa da vida nas 2 regiões é diferente
  - Não que esta diferença vai em uma direção ou outra
- Estamos conduzindo este teste ao nível de confiança de 99% ($\alpha = 0.01$)

Colecionar Dados

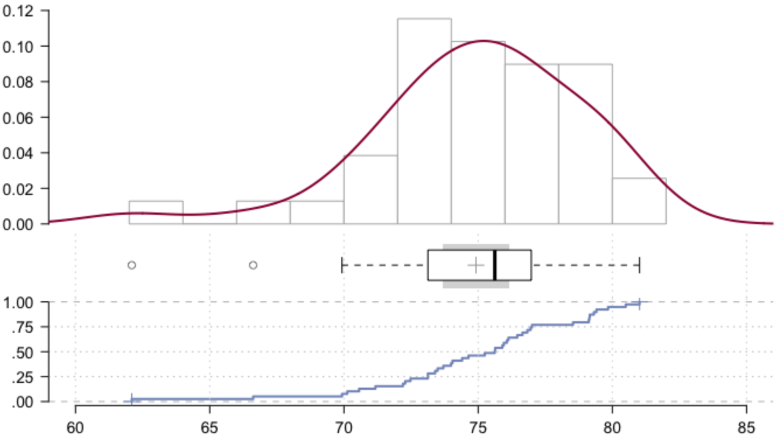
- Usamos os dados do arquivo "vidadados.RData"
- Derivado das bases de dados de Gapminder
- 197 países; 3 variáveis
    - `Pais`
    - `ExpVida`: Expectativa de Vida em Anos
    - `Regiao`: Região do mundo (para nos, "Amer" e "SSA")

Exploração dos Dados

```r
load("vidadados.RData")
amerSSA <- vidadados %>%
          filter(Regiao %in% c("Amer", "SSA"))
Desc(amerSSA$ExpVida[amerSSA$Regiao == "Amer"], plotit = FALSE)
```

```
## ------------------------------------------------------------------------------
## amerSSA$ExpVida[amerSSA$Regiao == "Amer"] (numeric)
##
##    length        n      NAs   unique       0s     mean   meanCI
##        39       39        0      = n        0  74.9192  73.6629
##             100.0%     0.0%              0.0%           76.1755
##
##       .05      .10      .25   median      .75      .90      .95
##   69.5961  70.4752  73.1265  75.6200  76.9795  79.3302  79.9050
##
##     range       sd    vcoef      mad      IQR     skew     kurt
##   18.9170   3.8755   0.0517   3.6961   3.8530  -0.9336   1.3853
##
## lowest : 62.095, 66.618, 69.927, 70.124, 70.563
## highest: 79.311, 79.407, 79.839, 80.499, 81.012
```

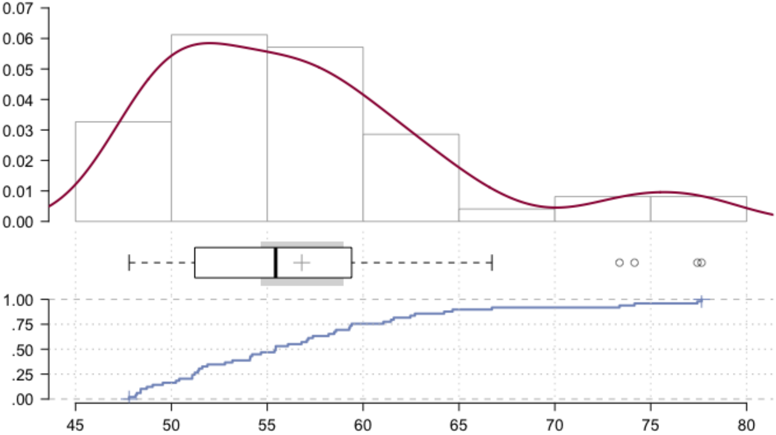
amerSSA$ExpVida[amerSSA$Regiao == "Amer"] (numeric)



```r
Desc(amerSSA$ExpVida[amerSSA$Regiao == "SSA"], plotit = FALSE)
```

```
## ------------------------------------------------------------------------------
## amerSSA$ExpVida[amerSSA$Regiao == "SSA"] (numeric)
##
##    length        n      NAs   unique       0s     mean   meanCI
##        49       49        0     = n         0  56.7985  54.6404
##             100.0%     0.0%              0.0%           58.9566
##
##       .05      .10      .25   median      .75      .90      .95
##   48.2764  48.6540  51.2190  55.4390  59.4000  65.0764  73.8428
##
##     range       sd    vcoef      mad      IQR     skew     kurt
##   29.8590   7.5133   0.1323   6.2566   8.1810   1.1464   0.8778
##
## lowest : 47.794, 48.132, 48.196, 48.397, 48.398
## highest: 66.718, 73.373, 74.156, 77.433, 77.653
```

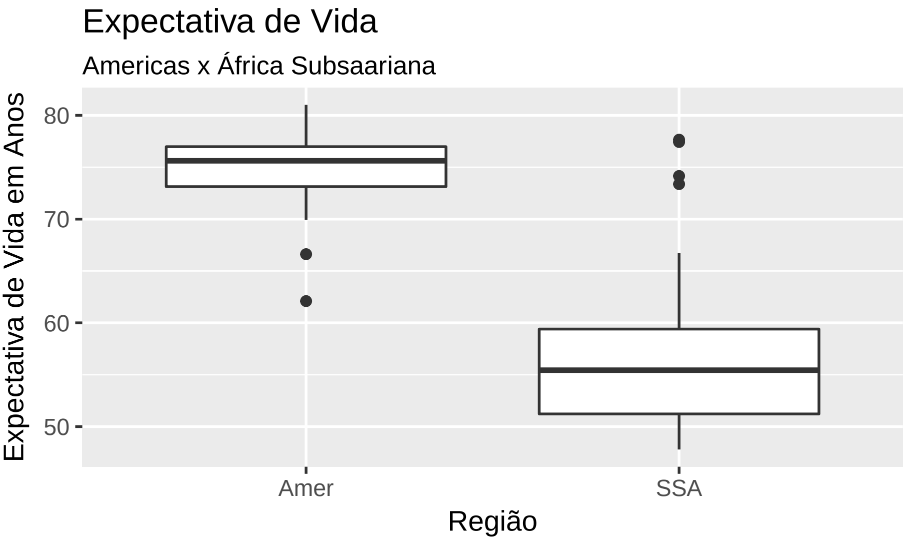
amerSSA$ExpVida[amerSSA$Regiao == "SSA"] (numeric)



```r
amerSSA %>% group_by(Regiao) %>%
            summarise_at(vars(ExpVida), funs(mean, sd))
```

```
## # A tibble: 2 × 3
##   Regiao     mean       sd
##   <fctr>    <dbl>    <dbl>
## 1   Amer 74.91921 3.875471
## 2    SSA 56.79851 7.513292
```

Boxplot das Regiões



Teste t das Regiões

```r
t.test(amerSSA$ExpVida[amerSSA$Regiao == "Amer"],
       amerSSA$ExpVida[amerSSA$Regiao == "SSA"],
       mu = 0, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  amerSSA$ExpVida[amerSSA$Regiao == "Amer"] and amerSSA$ExpVida[amerSSA$Reg
## t = 14.616, df = 74.885, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  15.65079 20.59060
## sample estimates:
## mean of x mean of y
##  74.91921  56.79851
```

Expectativa da Vida: Interpretação

- Rejeitamos H_0 que as duas regiões têm expectativas iguais
- Porque o teste foi de dois lados, só podemos dizer que não parecem iguais ao nível de 95%
- Precisa estudar mais para determinar o grau de diferença e porque existe

Lembrete: Qualidade dos Testes Estatísticas Dependem dos Números