AULAS 2-4: DISTRIBUIÇÕES DE PROBABILIDADE E TESTES DE HIPÓTESE

Análise Quantitativa de Dados Ambientais

Thiago S. F. Silva - tsfsilva@rc.unesp.br

August 28, 2015

Programa de Pós Graduação em Geografia - IGCE/UNESP

Distribuições de Probabilidade

Testes de Hipóteses

Erros Tipo I (α) e Tipo II (β)

DISTRIBUIÇÕES DE PROBABILIDADE

Definição Coloquial:

Um evento aleatório. A V.A. pode ser **discreta** (Ex. Captura), ou **contínua** (Ex. Temperatura)

Definição Matemática:

Uma função que associa um valor numérico a cada resultado possível de um experimento, dentro de um espaço amostral.

Mas que função é essa?

Definição Coloquial:

Um evento aleatório. A V.A. pode ser **discreta** (Ex. Captura), ou **contínua** (Ex. Temperatura)

Definição Matemática:

Uma função que associa um valor numérico a cada resultado possível de um experimento, dentro de um espaço amostral.

Mas que função é essa?

Uma função de **probabilidade**

Ex: Qual a probabilidade de uma planta carnívora capturar 5 a cada 10 insetos?

- Para variáveis aleatórias discretas, cada resultado tem uma probabilidade associada
- Então podemos dizer que a V.A. tem uma **função de massa de probabilidade (p.m.f.)**

Probabilide de Captura de Inseto:

$$f(x) = \begin{cases} p & \text{para } k = 1 \text{ (captura)} \\ q = (1 - p) & \text{para } k = 0 \text{ (sem captura)} \end{cases}$$

Distribuição de um único evento discreto, com probabilidade de sucesso p e probabilidade de falha $q = 1 - p$

É a distribuição de probabilidade discreta mais simples

Exemplos:

- Captura, Não-Captura
- Presença, Ausência
- Cara, Coroa
- Sim, Não
- Sucesso, Fracasso

Todas as distribuições de probabilidade são definidas por uma **função de probabilidade** e seus **parâmetros**.

Distribuição Bernoulli:

Parâmetros: $p, \quad (0 \leq p \leq 1, p \in \mathbb{R})$

f.m.p. (*p.m.f.*):

$$f(x) = P(x = k) = \begin{cases} q = (1 - p) & \text{para } k = 0 \\ p & \text{para } k = 1 \end{cases}$$

Exemplo: Probablidade de tirarmos 6 em um dado

$$D \sim \text{Bernoulli(p=1/6)}$$

Distribuição do número esperado de sucessos (k) em uma sequência de n realizações **independentes** de um evento discreto, com probabilidade de sucesso p e probabilidade de falha $q = 1 - p$

Distribuição Binomial: $X \sim B(n, p)$

Parâmetros: n ($n \in \mathbb{N}$); p ($0 \leq p \leq 1, p \in \mathbb{R}$)

Suporte: $x = k, k \in \{0, \ldots, n\}$

p.m.f:

$$f(x) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$\binom{n}{k}$ significa uma combinação de n resultados, k a k:

$\binom{n}{k} = C_k^n = \frac{n!}{k!(n-k)!}$

Exemplo:

Quantas combinações possíveis dos números de 1 a 4, 2 a 2?

$\binom{4}{2} = C_2^4 = \frac{4!}{2!(4-2)!} = \frac{4\times3\times2\times1}{2\times1\times(2\times1)} = \frac{24}{4} = 6$

```r
choose(4,2)

## [1] 6

combn(4,2)

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    1    2    2    3
## [2,]    2    3    4    3    4    4
```

p.m.f:

$$f(x) = \binom{n}{k} p^k (1-p)^{n-k}$$

Entendendo:

Queremos k sucessos com probabilidade $p^k = P(C_1 \cap C_2 \ldots \cap C_k)$

Se temos p sucessos, temos necessariamente $n - k$ falhas, com probabilidade $q = (1-p)^{n-k}$

Como a ordem não importa, então existem $\binom{n}{k}$ maneiras de se obter essa combinação de sucessos e fracassos ao longo de n realizações.

Exemplo: Qual a probabilidade de 7 insetos serem capturados, depois de 10 visitas, se a probabilidade de captura é de 0.2 por inseto?

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \times p^k \times (1-p)^{n-k}$$

$$P(x = 7) = \binom{10}{7} \times 0.2^7 \times 0.8^3 = \frac{10!}{7!(3)!} \times 0.2^7 \times 0.8^3$$

$$P(x = 7) = 120 \times 1.3 \times 10^{-5} \times 0.512$$

$$P(x = 7) = 7.9 \times 10^{-4}$$

Exercício 1: Qual a probabilidade de eu jogar uma moeda 5 vezes e obter 3 caras?

$$P(x=k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} \times p^k \times (1-p)^{n-k}$$

Exercício 1: Qual a probabilidade de eu jogar uma moeda 5 vezes e obter 3 caras?

$P(x = 3) =?$, $X \sim B(p = 0.5, n = 5)$

$$P(x = 3) = \binom{5}{3} \times 0.5^3 \times 0.5^2 = \frac{5!}{3!(5-3)!} \times 0.5^3 \times 0.3^3$$

$$P(x = 3) = 10 \times 0.125 \times 0.25$$

$$P(x = 3) = 0.3125$$

Exercício 1: Qual a probabilidade de eu jogar uma moeda 5 vezes e obter 3 caras?

```
p_x <- choose(5,3) * 0.5^3 * 0.5^2

p_x

## [1] 0.3125

dbinom(3,size=5,prob=0.5)

## [1] 0.3125
```

Exercício 2: Qual a probabilidade de eu jogar uma moeda 5 vezes e obter entre 2 e 4 caras?

Podemos pensar no problema como a união de três probabilidades:

$P(x = 2) \cup P(x = 3) \cup p(x = 4)$

Exercício 2: Qual a probabilidade de eu jogar uma moeda 5 vezes e obter entre 2 e 4 caras?

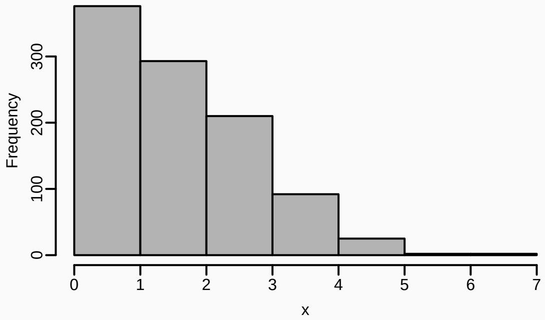
Podemos pensar no problema como a união de três probabilidades:

$$P(x = 2) \cup P(x = 3) \cup p(x = 4)$$

```
p2 <- dbinom(2,size=10, prob=0.2)
p3 <- dbinom(3,size=10, prob=0.2)
p4 <- dbinom(4,size=10, prob=0.2)

p2 + p3 + p4

## [1] 0.5913969
```

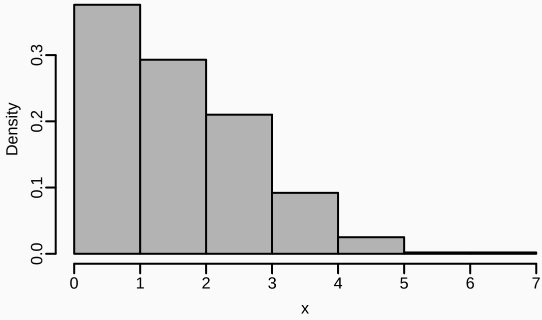
Podemos também simular diferentes réplicas, e observar a frequência dos resultados:

```
set.seed(40) #"fixa" a geração do número aleatório

x <- rbinom(1000,size=10,prob=0.2)

hist(x,breaks=7,main=NA,col="gray70")
```



Podemos também simular diferentes réplicas, e observar a frequência dos resultados:

```
c(p2,p3,p4); p2 + p3 + p4

## [1] 0.30198989 0.20132659 0.08808038
## [1] 0.5913969

hist(x,breaks=7,main=NA,col="gray70",freq=F)
```



POR QUE USAR DISTRIBUIÇÕES?

- Qual a vantagem de conhecermos uma distribuição de probabilidade?

- Você espera que a maioria dos dados na natureza siga uma distribuição específica?

As distribuições possuem **propriedades conhecidas**. Para a Distribuição Binomial:

Esperança (Primeiro Momento): O valor ”esperado” de uma V.A.

$$E[X] = \sum_{i=1}^{n} x_i p_i$$

As distribuições possuem **propriedades conhecidas**. Para a Distribuição Binomial:

Esperança (Primeiro Momento): O valor ”esperado” de uma V.A.

$$E[X] = \sum_{i=1}^{n} x_i p_i$$

Variância (Segundo Momento): A dispersão de uma V.A.

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

Distribuição Binomial: $X \sim B(n, p)$

Parâmetros: n ($n \in \mathbb{N}$); p ($0 < p < 1, p \in \mathbb{R}$)

Suporte: $x = k, k \in \{0, \ldots, n\}$

p.m.f:

$$f(x) = \binom{n}{k} p^k (1-p)^{n-k}$$

$E[X] = np$

$Var[X] = np(1-p)$

Exemplo: Se a probabilidade de captura é 0.2, qual o número médio de capturas eu espero obter após 10 visitas?

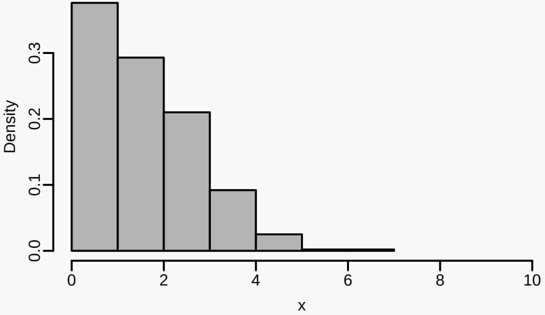
$$E[x] = np = 10 \times 0.2 = 2 \text{ capturas}$$

Exemplo: Como varia o número de capturas após 10 visitas?

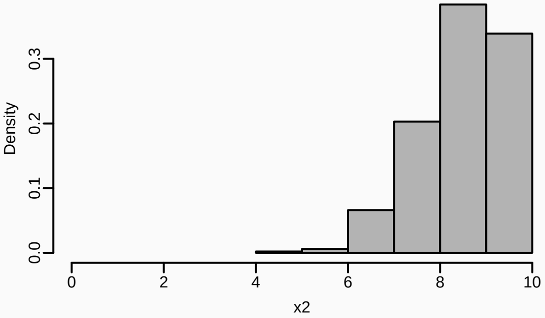
$$Var[X] = np(1-p) = 10 \times 0.2 \times (1-0.2) = 10 \times 0.2 \times 0.8 = 1.6$$

DISTRIBUIÇÃO BINOMIAL

$$E[X] = 2$$
$$Var[X] = 1.6$$



$n = 10, p = 0.9$
$E[X] = 9$
$Var[X] = 0.9$



- **Bernoulli** (Binomial com n=1): Sucesso em um evento.
- **Binomial** ($X \sim B(n, p)$): Sucesso em eventos sucessivos.
- **Multinomial** ($X \sim M(n, p_1, \ldots, p_k)$): Generalização da binomial para mais de dois resultados possíveis.
- **Poisson** ($X \sim Pois(\lambda)$): Sucesso em um número desconhecido de eventos.
- **Binomial Negativa** ($X \sim NB(r, p)$): Número de falhas acumuladas até que um certo número de sucessos ocorra.
- **Geométrica** ($X \sim Geom(p)$): Número de realizações até ocorrer uma falha.
- **Beta-binomial** ($X \sim BetaBin(p, \theta)$): Binomial negativa com probabilidade de sucesso variável.

Poisson: A Binomial modela o número de sucessos esperados com um numero fixo de realizações. A distribuição Poisson modela a ocorrência de sucessos em situações onde o número de realizações é infinito. O exemplo mais comum são dados de contagem de indivíduos em parcelas, ou ao longo de um intervalo de tempo.

Distribuição Poisson: $X \sim Pois(\lambda)$

Parâmetros: λ ($\lambda > 0$)

Suporte: $x = k, k \in \{0, \dots, n\}$

p.m.f:

$$f(x) = \frac{lambda^k}{k!} e^{-\lambda}$$

$E[X] = \lambda$

$Var[X] = \lambda$

Ex.: Se em média eu observo 50 indivíduos por parcela, qual a probabilidade de eu observar uma parcela com 100 indivíduos?

$$f(x) = \frac{lambda^k}{k!} e^{-\lambda}$$

$$P(x = 100) = \frac{50^{100}}{100!} \times e^{-50}$$

$$P(x = 100) = 845272575844 \times 1.92875 \times 10^{-22}$$

$$P(x = 100) = 1.630319 \times 10^{-10}$$

```
dpois(100,50)

## [1] 1.630319e-10
```

- Até agora falamos de V.A. discretas
- Mas e se os dados que queremos modelar são contínuos?
- **Exemplo:** Qual a probabilidade da temperatura máxima de hoje ser 32°C?

- Uma V.A. contínua pode assumir infinitos valores
- Se assumimos que: $P \approx F = \frac{n_i}{N}$
- Qual dos dois resultados tem probabilidade maior?

P(temperatura máxima de hoje) = 32°C?

ou

P(temperatura máxima de hoje) = 32.354321°C?

Qual dos dois resultados tem probabilidade maior?

P(temperatura máxima de hoje) = 32°C?

ou

P(temperatura máxima de hoje) = 32.354321°C?

Os dois tem a mesma probabilidade, que é **zero**.

$$P(n_i) \approx F = \frac{n_i}{N} = \frac{1}{\inf} = 0$$

$$\lim_{N \to \inf} P(x) = 0$$

Para V.A. contínuas, ao invés de massas de probabilidade, falamos de **densidades de probabilidade**, dentro de um intervalo de valores.

As distribuições de probabilidade contínuas tem, desta maneira, **funções de densidade de probabilidade (f.d.p ou *p.d.f*)**

Qual a probabilidade da temperatura máxima de hoje estar entre 32 e 33°C?

Qual a probabilidade da temperatura máxima de hoje ser maior que 32°C?

Qual a distribuição contínua mais utilizada?

Qual a distribuição contínua mais utilizada?

Distribuição Normal (Gaussiana): $X \sim N(\mu, \sigma)$

Parâmetros: μ ($\mu \in \mathbb{R}$)); σ ($\sigma > 0$)

Suporte: $X \in \mathbb{R}$

p.d.f:

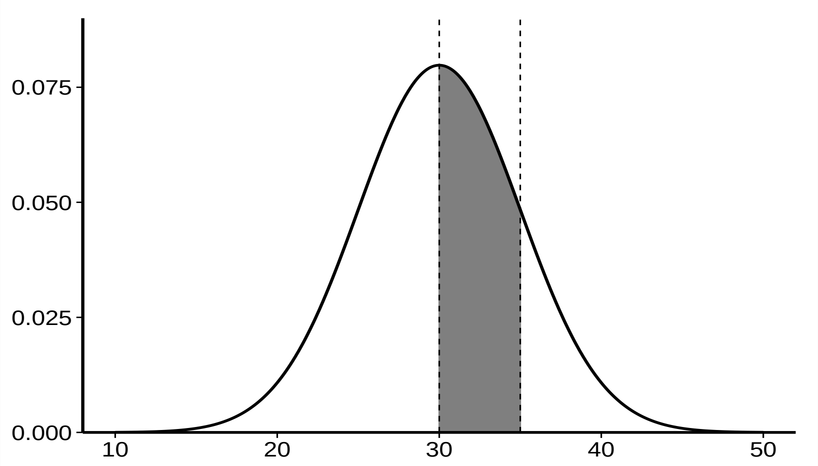
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$E(X) = \mu$

$Var(X) = \sigma^2$ (Muitas vezes usamos o desvio padrão: $\sqrt{\sigma^2} = \sigma$)

Exemplo: Qual a probabilidade de observamos uma temperatura entre 30°C e 35°C, se a média histórica é $\mu = 30$, e o desvio padrão é $\sigma = 5$?

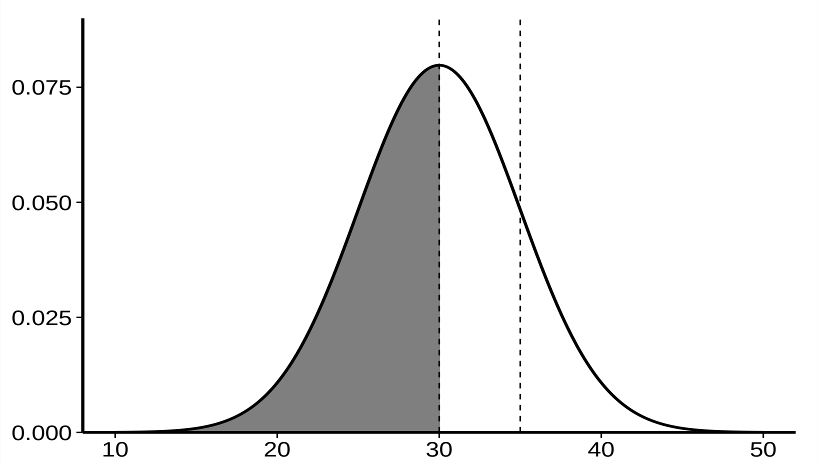
DISTRIBUIÇÃO NORMAL - EXEMPLO



DISTRIBUIÇÃO NORMAL - EXEMPLO

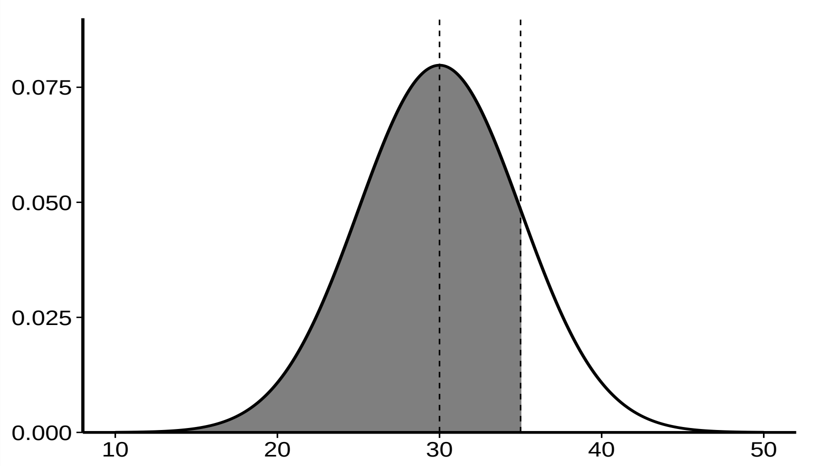
```
# No R
media <- 30
desvio <- 5
tmin <- 30
tmax <- 35

# A maneira mais fácil de calcular uma probabilidade é de maneira cumulativa: P(x <= 30)
#O comando ”pnorm” faz isso:

p.tmin <- pnorm(tmin,mean=media,sd=desvio)

p.tmin

## [1] 0.5
```

DISTRIBUIÇÃO NORMAL - EXEMPLO



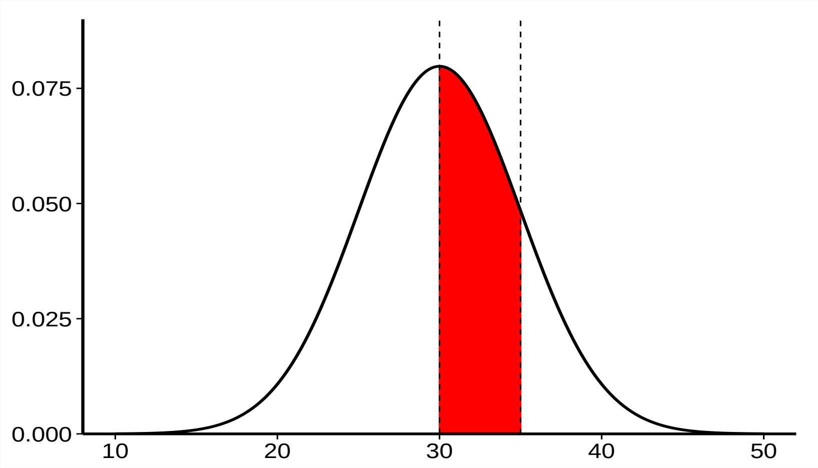
DISTRIBUIÇÃO NORMAL - EXEMPLO

```r
# No R
media <- 30
desvio <- 5
tmin <- 30
tmax <- 35

# calculamos também a probabilidade cumulativa de x <= 35

p.tmax <- pnorm(tmax,mean=media,sd=desvio)

p.tmax

## [1] 0.8413447
```


0.075
0.050
0.025
0.000
10
20
30
40
50

DISTRIBUIÇÃO NORMAL - EXEMPLO

```
# Sabendo as duas probabilidades cumulativas, é só subtrair

p.tmax - p.tmin

## [1] 0.3413447
```


0.075
0.050
0.025
0.000
10
20
30
40
50

- Normal: $X \sim N(\mu, \sigma)$
- Gamma: $X \sim Gamma(s, a)$
- Exponencial: $X \sim Exp(\lambda)$
- Beta: $X \sim Beta(a, b)$
- Lognormal: $X \sim logN(\mu, \sigma)$
- Qui-quadrado: $X \sim \chi^2(df)$

TESTES DE HIPÓTESES

O mecanismo dos testes de hipóteses **paramétricos** seguem sempre a mesma lógica:

1. Formule uma hipótese quantitativa

O mecanismo dos testes de hipóteses **paramétricos** seguem sempre a mesma lógica:

1. Formule uma hipótese quantitativa
2. Defina uma estatística de interesse que descreva essa quantidade

O mecanismo dos testes de hipóteses **paramétricos** seguem sempre a mesma lógica:

1. Formule uma hipótese quantitativa
2. Defina uma estatística de interesse que descreva essa quantidade
3. Assuma uma distribuição para esta estatística de interesse, caso a hipótese seja verdadeira
4. Especifique os parâmetros dessa distribuição

O mecanismo dos testes de hipóteses **paramétricos** seguem sempre a mesma lógica:

1. Formule uma hipótese quantitativa
2. Defina uma estatística de interesse que descreva essa quantidade
3. Assuma uma distribuição para esta estatística de interesse, caso a hipótese seja verdadeira
4. Especifique os parâmetros dessa distribuição
5. Calcule a probabilidade de se obter a estatística de interesse observada, **ou uma mais extrema**, a partir da sua amostra, se a sua hipótese for verdadeira.

O mecanismo dos testes de hipóteses **paramétricos** seguem sempre a mesma lógica:

1. Formule uma hipótese quantitativa
2. Defina uma estatística de interesse que descreva essa quantidade
3. Assuma uma distribuição para esta estatística de interesse, caso a hipótese seja verdadeira
4. Especifique os parâmetros dessa distribuição
5. Calcule a probabilidade de se obter a estatística de interesse observada, **ou uma mais extrema**, a partir da sua amostra, se a sua hipótese for verdadeira.
6. Avalie a ”força da evidência” em relação sua hipótese, com base na probabilidade observada para a amostra.

O mecanismo dos testes de hipóteses **paramétricos** seguem sempre a mesma lógica:

1. Formule uma hipótese quantitativa
2. Defina uma estatística de interesse que descreva essa quantidade
3. Assuma uma distribuição para esta estatística de interesse, caso a hipótese seja verdadeira
4. Especifique os parâmetros dessa distribuição
5. Calcule a probabilidade de se obter a estatística de interesse observada, **ou uma mais extrema**, a partir da sua amostra, se a sua hipótese for verdadeira.
6. Avalie a ”força da evidência” em relação sua hipótese, com base na probabilidade observada para a amostra.

Pergunta: Estou comparando duas amostras pareadas (X, Y), com N observações, e quero saber se existe diferença entre elas.

Hipótese: Se não há diferença, $P(x_i > y_i) = P(y_i > x_i) = 0.5$. Podemos pensar em $x_i > y_i$ como um sucesso, e $y_i > x_i$ como um fracasso

Estatística de interesse - W: quantas vezes observamos $x_i > y_i$?

Distribuição de W?

Pergunta: Estou comparando duas amostras pareadas (X, Y), com N observações, e quero saber se existe diferença entre elas.

Hipótese: Se não há diferença, $P(x_i > y_i) = P(y_i > x_i) = 0.5$. Podemos pensar em $x_i > y_i$ como um sucesso, e $y_i > x_i$ como um fracasso

Estatística de interesse - W: quantas vezes observamos $x_i > y_i$?

Distribuição de W: $W\, Bin(n = N, p = 0.5)$

Qual a probabilidade de obtermos o valor observado de W ou maior, na nossa amostra, se $W\, Bin(n = N, p = 0.5)$?

```
x <- c(0,3,6,5,3,7,8,9,2,4,6,7)
y <- c(3,4,5,6,9,7,1,2,3,4,5,6)

n=length(x); n

## [1] 12

d <- x-y; d

##  [1] -3 -1  1 -1 -6  0  7  7 -1  0  1  1

w <- sum(x-y > 0); w

## [1] 5

dbinom(w,size=12,prob=0.5)

## [1] 0.1933594
```

EXEMPLO:TESTE DOS SINAIS

```r
# Mas eu quero saber P(w >= 5)

probs <-vector(8,mode='numeric')

for(i in c(5:12)){
  w=i
  probs[i] <- dbinom(w,size=12,prob=0.5)
}

probs

##  [1] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.1933593750
##  [6] 0.2255859375 0.1933593750 0.1208496094 0.0537109375 0.0161132813
## [11] 0.0029296875 0.0002441406

sum(probs)

## [1] 0.8061523
```

EXEMPLO:TESTE DOS SINAIS

```r
binom.test(5,12,p=0.5,alternative="greater")

##
## ^^IExact binomial test
##
## data:  5 and 12
## number of successes = 5, number of trials = 12, p-value = 0.8062
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.1810248 1.0000000
## sample estimates:
## probability of success
##              0.4166667
```

Teste χ^2 (chi-quadrado, chi pronuncia-se ”qui”)

Pergunta: Contei os indivíduos em 3 habitats: $N_F = 86$,$N_P = 3$,$N_A = 11$. Cada habitat estava representado na seguinte proporção: Floresta(75%), Pastagem (10%), Agricultura(15%) . Será que existe uma preferência dos indivíduos por um determinado habitat?

Hipótese: Se não há preferência, a quantidade esperada de indivíduos em cada habitat só é afetada pela proporção de cada um. Se a quantidade observada for diferente da esperada, há indício de preferência.

Estatística: $X^2 = \sum \frac{(O-E)^2}{E}$

Distribuição de X^2: $X^2 \sim \chi^2(k)$ (k = graus de liberdade = $N - 1$)

Qual a probabilidade de obtermos o valor observado de X^2 ou mais extremo, na nossa amostra?

```r
obs <- c(86,3,11)
habs <- c(0.75,0.1,0.15)
esp <- rep(sum(obs),3) * habs

x2 <- sum((obs-esp)^2/(esp))
x2

## [1] 7.58

dchisq(x2,df=2)

## [1] 0.0112978

# A distribuição qui-quadrada é contínua, então não dá pra somar. Mas existe uma função cumulativa
1-pchisq(x2,df=2)

## [1] 0.0225956

chisq.test(obs,p=habs)

##
## ^^IChi-squared test for given probabilities
##
## data:  obs
## X-squared = 7.58, df = 2, p-value = 0.0226
```

Pergunta

Um reservatório com concentrações de clorofila maiores do que 3 mg.m^{-3} pode ser considerado eutrófico. Eu coleto 20 amostras de água e determino uma concentração média de 2.55 mg.m^{-3}, com um desvio padrão de 0.9 mg.m^{-3}. Será que meu reservatório é eutrófico?

H$_0$: O reservatório está contaminado, então $\mu = 3$

H$_1$: O reservatório não está contaminado, então $\mu < 3$

$P(\bar{X} >= 2.55|\mu = 3)$?

Estatística: Z

$$Z = \frac{\bar{X} - \mu}{E.P.} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Distribuição de Z?

H$_0$: O reservatório está contaminado, então $\mu = 3$

H$_1$: O reservatório não está contaminado, então $\mu < 3$

$P(\bar{X} >= 2.55|\mu = 3)$?

Estatística: Z

$$Z = \frac{\bar{X} - \mu}{E.P.} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Distribuição de Z: $Z \sim N(0, 1)$

Uma confusão comum é confundir desvio padrão (*standard deviation*) e erro padrão (*standard error*). Qual a diferença?

Uma confusão comum é confundir desvio padrão (*standard deviation*) e erro padrão (*standard error*). Qual a diferença?

O desvio padrão mede a dispersão dos dados observados.

A partir desses dados, podemos calcular a média ($\bar{X}$), um estimador da média da população (μ). Se tomarmos amostras diferentes, teremos $\bar{X}$ diferentes. Qual a distribuição de $\bar{X}$?

Uma confusão comum é confundir desvio padrão (*standard deviation*) e erro padrão (*standard error*). Qual a diferença?

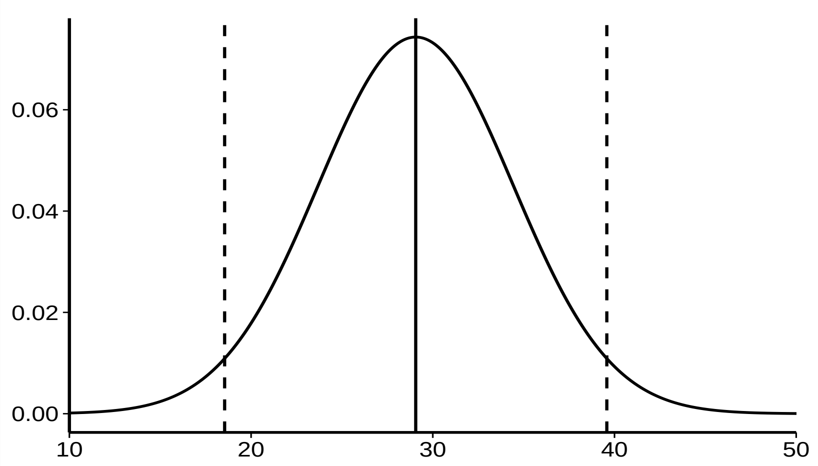
O desvio padrão mede a dispersão dos dados observados.

A partir desses dados, podemos calcular a média ($\bar{X}$), um estimador da média da população (μ). Se tomarmos amostras diferentes, teremos $\bar{X}$ diferentes. Qual a distribuição de $\bar{X}$?

O erro padrão mede a dispersão esperada (desvio padrão) de $\bar{X}$, não dos dados originais

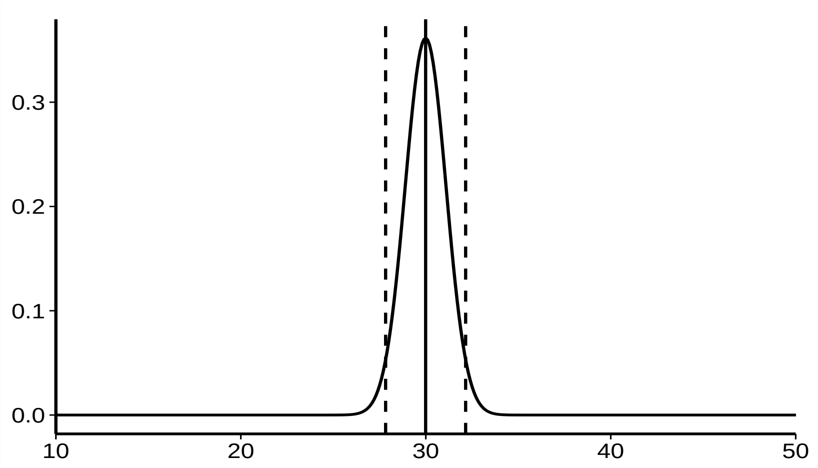
Erro Padrão da Média: $E.P. = \frac{\sigma}{\sqrt{n}}$

```
# Tomamos uma amostra aleatória com X ~ N(30,5) e n=50
set.seed(20)
x <- rnorm(20,30,5); x

##  [1] 35.81343 27.07038 38.92733 23.33703 27.76717 32.84803 15.55141 25.65491
##  [9] 27.69149 27.22230 29.89932 29.24809 26.85937 36.61610 22.39325 27.81286
## [17] 34.85289 30.14111 29.57109 31.94607

# Calculamos a média e o desvio padrão
x_barra <- mean(x); x_barra

## [1] 29.06118

s <- sd(x); s

## [1] 5.365711
```

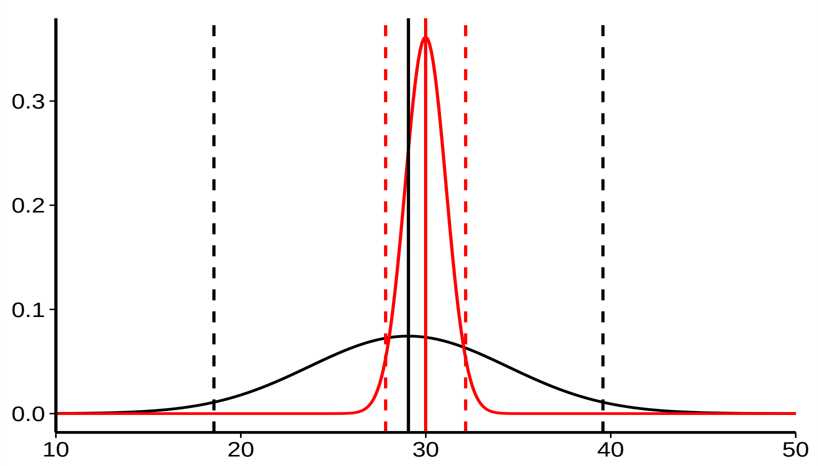

0.06
0.04
0.02
0.00
10
20
30
40
50

DESVIO PADRÃO VS. ERRO PADRÃO

```r
# Mas podemos repetir essa amostragem 10000 vezes, e ter 10000 médias diferentes
medias <- vector(10000,mode='numeric')

for (i in c(1:10000)){
  x <- rnorm(20,30,5)
  medias[i] <- mean(x)
}

mean(medias)

## [1] 29.99082

sd(medias)

## [1] 1.103708
```


0.3
0.2
0.1
0.0
10
20
30
40
50

DESVIO PADRÃO VS. ERRO PADRÃO



```r
# De fato

sd(medias)

## [1] 1.103708

5/sqrt(20)

## [1] 1.118034
```

H_0: O reservatório está contaminado, então $\mu = 3$

H_1: O reservatório não está contaminado, então $\mu < 3$

$P(\bar{X} >= 2.55|\mu = 3)$?

Estatística: Z

$$Z = \frac{\bar{X} - \mu}{E.P.} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Distribuição de Z?

H_0: O reservatório está contaminado, então $\mu = 3$

H_1: O reservatório não está contaminado, então $\mu < 3$

$P(\bar{X} >= 2.55 | \mu = 3)$?

Estatística: Z

$$Z = \frac{\bar{X} - \mu}{E.P.} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

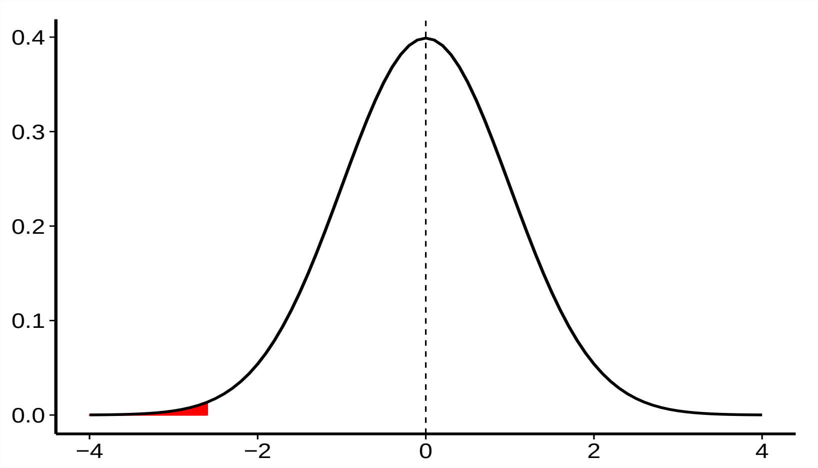
Distribuição de Z: $Z \sim N(0, 1)$

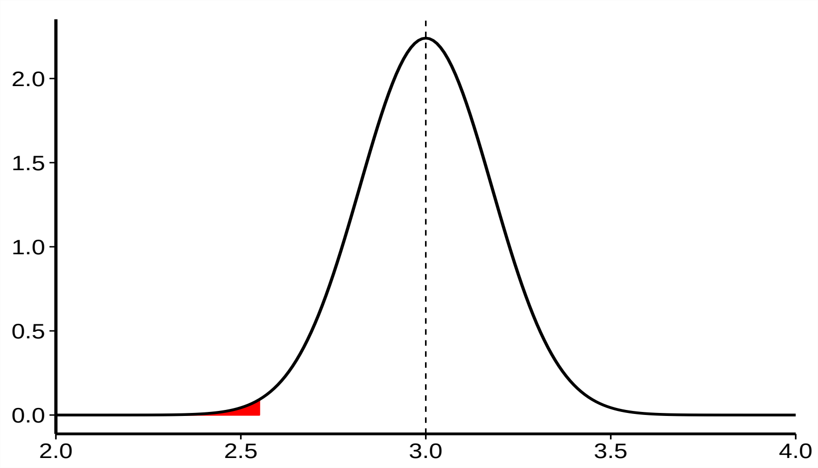
```
n = 20
x <-  c(1.85, 2.64, 3.63, 1.94, 2.41, 2.74, 2.85, 3.07, 1.29,
        1.50, 1.55, 1.69, 3.70, 3.25, 2.47, 1.95, 3.33, 2.21, 3.02, 3.98)
x_barra <- mean(x)
x_barra

## [1] 2.5535

s <- sd(x)
s

## [1] 0.7964016

mu <- 3

x_barra-mu

## [1] -0.4465

(x_barra-mu)/(s/sqrt(n))

## [1] -2.507289
```

EXEMPLO: TESTE Z PARA A MÉDIA

```
z <- (x_barra-mu)/(s/sqrt(n))

pnorm(z,0,1)

## [1] 0.006083066

# Ou simplesmente

pnorm(x_barra,mu,s/sqrt(n))

## [1] 0.006083066
```

TESTE Z - VISUALMENTE




2.0
1.5
1.0
0.5
0.0
2.0
2.5
3.0
3.5
4.0

Pergunta

Estou interessado em saber se a concentração de clorofila varia entre dois reservatórios.

O primeiro tem um concentração de 2.5 mg.m^{-3}, e o segundo de 2.8 mg.m^{-3} (diferença de 0.3 mg.m^{-3}). Qual a evidência de que as concentrações são diferentes?

H_0: Os reservatórios são iguais, então $\mu_1 = \mu_2$, ou $\mu_1 - \mu_2 = 0$

H_1: Os reservatórios são diferentes, então $\mu_1 \neq \mu_2$, ou $\mu_1 - \mu_2 \neq 0$

$P(\bar{X}_1 - \bar{X}_1 >= 0.3 | \mu_1 - \mu_2 = 0)$?

Estatística: Z

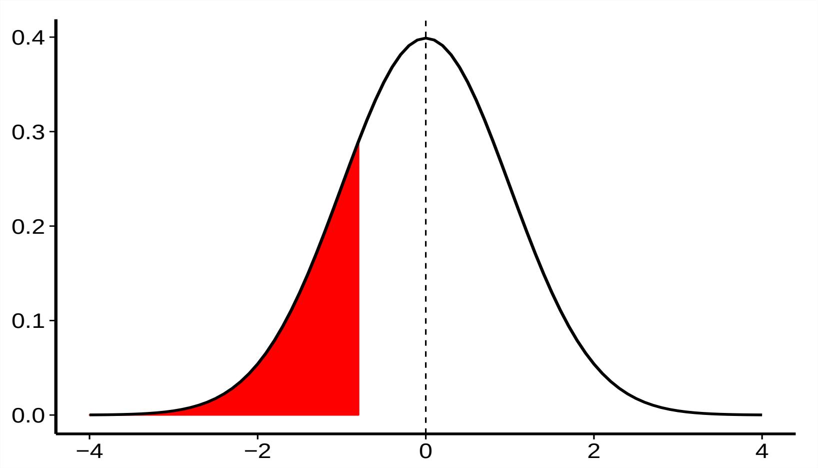
$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{E.P._1^2 + E.P._2^2}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}}}$$

```r
n=20
x1 <-  c(1.85, 2.64, 3.63, 1.94, 2.41, 2.74, 2.85, 3.07, 1.29,1.50, 1.55, 1.69,
         3.70, 3.25, 2.47, 1.95, 3.33, 2.21, 3.02, 3.98)

x2 <- c(2.79, 2.61, 3.72, 1.58, 2.02, 2.38, 2.70, 3.18, 3.96, 1.53, 1.51, 2.08,
        3.34, 3.76, 2.61, 3.98, 3.50, 2.34, 2.95, 3.91)

x_barra1 = mean(x1); x_barra2 = mean(x2)

x_barra1; x_barra2

## [1] 2.5535
## [1] 2.8225

s1 = sd(x1); s2=sd(x2)
s1;s2

## [1] 0.7964016
## [1] 0.8284092

mu1 = mu2 = 0
```

EXEMPLO: TESTE Z PARA A MÉDIA

```r
z <- ((x_barra1-x_barra2)-(mu1-mu2))/(sqrt(s1^2/n)+sqrt(s2^2/n))
z

## [1] -0.7403967

pnorm(z,0,1)

## [1] 0.2295297

# Ou simplesmente

pnorm(x_barra1-x_barra2,mu1-mu2,sqrt(s1^2/n)+sqrt(s2^2/n))

## [1] 0.2295297
```

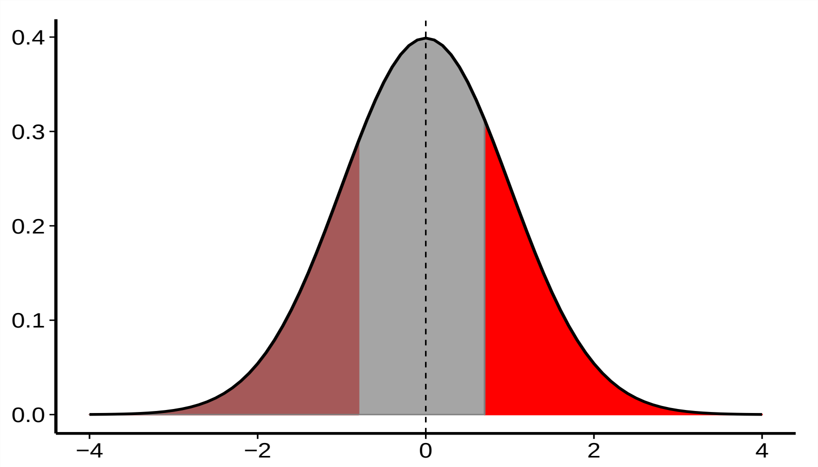
TESTE Z - VISUALMENTE



Mas isso é só metade da história…a princípio, não sabemos na realidade qual lago é maior e qual é menor, então temos que considerar tanto que $\mu_1 > \mu_2$ quanto $\mu_2 > \mu_1$.

```r
z_min <- ((x_barra1-x_barra2)-(mu1-mu2))/(sqrt(s1^2/n)+sqrt(s2^2/n))
z_max <- ((x_barra2-x_barra1)-(mu2-mu1))/(sqrt(s2^2/n)+sqrt(s1^2/n))

pnorm(z_min,0,1)

## [1] 0.2295297

pnorm(z_max,0,1)

## [1] 0.7704703

p_final <- pnorm(z_min,0,1) + (1 - pnorm(z_max,0,1))

p_final

## [1] 0.4590593
```

TESTE Z - VISUALMENTE

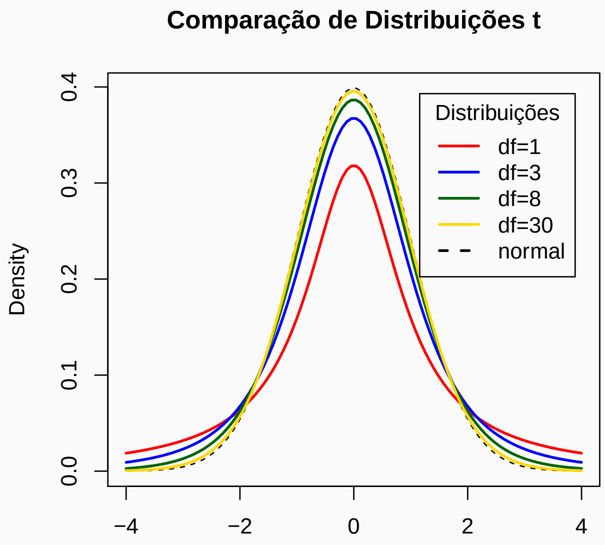


- Até agora, assumimos que o desvio padrão da amostra aproxima o desvio padrão da população. Mas quanto menor a amostra, menos isso é verdade.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \neq \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

- O uso de s em vez de σ introduz um erro, ou o que chamamos de um **viés (bias)**
- Felizmente, existe uma maneira simples de corrigir esse viés:
- A distribuição **t de Student**

- Student = William Gosset
- Pelo Teorema do Limite Central, s aproxima σ para "grandes" amostras ($n \geq 30$). Mas, se as amostras são pequenas, isso não vale.
- Para pequenas amostras, essa estatística se aproxima mais de uma distribuição t de Student.
- O único parâmetro de t é $\nu = (n - 1)$, onde n é o número de observações. Esse parametro é conhecido como "graus de liberdade".
- Quando ($n \geq 30$), t se aproxima de uma distribuição normal.



Comparação de Distribuições t
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
Distribuições
df=1
df=3
df=8
df=30
normal

CORRIGINDO NOSSOS EXEMPLOS ANTERIORES

```
t <- (x_barra-mu)/(s/sqrt(n))

pnorm(t,0,1);pt(t,n-1) # neste caso, chamamos a estatística de t, e não z

## [1] 0.006083066
## [1] 0.01070439

# ou, usando o comando interno do R
t.test(x,mu=3,alternative='less')

##
## ^^IOne Sample t-test
##
## data:  x
## t = -2.5073, df = 19, p-value = 0.0107
## alternative hypothesis: true mean is less than 3
## 95 percent confidence interval:
##      -Inf 2.861425
## sample estimates:
## mean of x
##    2.5535
```

CORRIGINDO NOSSOS EXEMPLOS ANTERIORES

```r
t <- ((x_barra1-x_barra2)-(mu1-mu2))/sqrt((s1^2/n1)+(s2^2/n2))

pnorm(t,0,1); pt(t,n-1)

## [1] 0.1475784
## [1] 0.1541463

t.test(x1,x2, alternative="less")

##
## ^^IWelch Two Sample t-test
##
## data:  x1 and x2
## t = -1.0469, df = 37.941, p-value = 0.1509
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf 0.1642312
## sample estimates:
## mean of x mean of y
##    2.5535    2.8225
```

Teste t para amostras independentes e variância conhecida:

$$\bar{D} = \bar{X} - \bar{Y} \qquad Var(\bar{D}) = \sigma_1^2/n_1 + \sigma_2^2/n_2 \qquad \frac{\bar{D}-\mu_D}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1)$$

Teste t para amostras independentes e variância desconhecidas e iguais:

$$\bar{D} = \bar{X} - \bar{Y} \qquad S_c^2 = \frac{(n_1-1)S_X^2+(n_2-1)S_Y^2}{(n_1-1)+(n_2-1)} \qquad \frac{\bar{D}-\mu_D}{\sqrt{S_c^2(1/n_1+1/n_2)}} \sim t_{(n_1+n_2-2)}$$

Teste t para amostras independentes e variância desconhecidas e diferentes:

$$\bar{D} = \bar{X} - \bar{Y} \quad \hat{S}^2 = S_X^2/n_1 + S_Y^2/n_2 \quad \frac{\bar{D} - \mu_D}{\sqrt{S_X^2/n_1 + S_Y^2/n_2)}} \sim t_{(\nu)}$$

Teste t para amostras pareadas

$$\bar{D} = \frac{\sum D_i}{N} \quad S_D^2 = \frac{1}{n-1} \sum D_i - \bar{D} \quad \frac{\bar{D} - \mu_D}{\sqrt{S_D^2/n}} \sim t_{(n-1)}$$

ERROS TIPO I (α) E TIPO II (β)

LEMBRETE IMPORTANTE: o teste é sempre baseado na distribuição amostral da estatística de interesse, e não na distribuição dos dados originais!

Imaginemos um teste z para comparação das médias de duas amostras:

$X_1 \sim N(\mu = 10, \sigma = 5)$ e $X_2 \sim N(\mu = 12, \sigma = 5)$.

Cada população foi amostrada com $n = 30$.

```
set.seed(1979)
x1 <- rnorm(30, 10, 5)
x2 <- rnorm(30, 12, 5)
```

Como sabemos, $\bar{X}$ e s aproximam (estimam) μ e σ. Quanto maior o n, melhor a estimação.

```
mean(x1)
## [1] 9.413282
mean(x2)
## [1] 12.40707
```

```
sd(x1)
## [1] 5.012927
sd(x2)
## [1] 3.540709
```

Poderíamos formular duas hipóteses

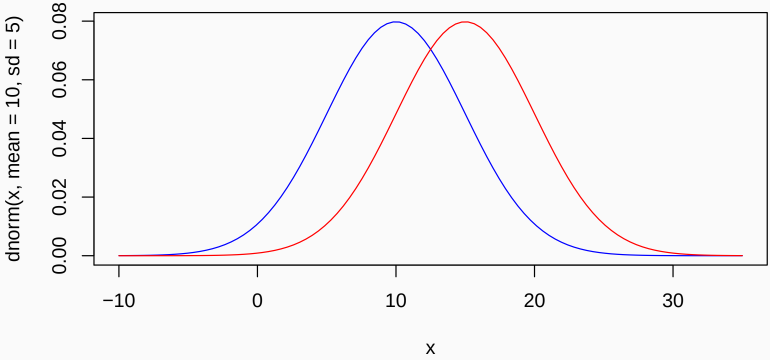
H_1: $\mu_1 - \mu_2 = 0$

H_2: $\mu_1 - \mu_2 = 2$

Neste caso, a nossa H_2 corresponde exatamente à realidade, mas só para fins didáticos.

Podemos até visualizar como X_1 e X_2 estão distribuídos:

```
curve(dnorm(x, mean = 10, sd = 5), -10, 35, , col = "blue", cex = 2)
curve(dnorm(x, mean = 15, sd = 5), -10, 35, , col = "red", add = T)
```



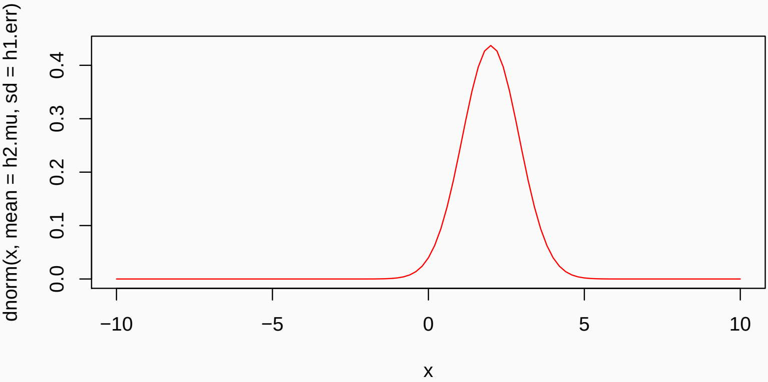
Mas o nosso teste se baseia na distribuição amostral da estatística ($\bar{X}_D - \mu_D$), e não das variáveis ($\bar{X}_1$ e $\bar{X}_2$).

$$\text{Lembrando: } \sigma_\mu = \frac{\sigma}{\sqrt{N}}$$

```
h1.mu <- 0

h2.mu <- 2

h1.err <- h2.err <- 5/sqrt(30)
```

ERROS TIPO I (α)

```
curve(dnorm(x, mean = h2.mu, sd = h1.err), -10, 10, , col = "red")
```




```
curve(dnorm(x, mean = h1.mu, sd = h0.err), -10, 10, , col = "blue", add = T)

## Error in dnorm(x, mean = h1.mu, sd = h0.err): object 'h0.err' not found
```

O teste de hipótese clássico (”ritual nulo”), como vimos, só se baseia em refutar uma das hipóteses, independente de outras hipóteses.

1. Definimos a estatística de interesse: $(\bar{X}_D - \mu_D)/E.P.$ $(H_1 : \mu_D = 0)$
2. Estimamos a estatística com base na amostra
3. Assumimos uma distribuição para essa estatística (z)
4. Estabelecemos nosso nível de significância ($\alpha = 5\%$)
5. Calculamos a probabilidade de observarmos uma estatística z maior ou igual ao valor crítico ($P(z \geq z_{0.05}$)
6. Com base nesse probabilidade, rejeitamos ou não a hipótese

Para o nosso caso:

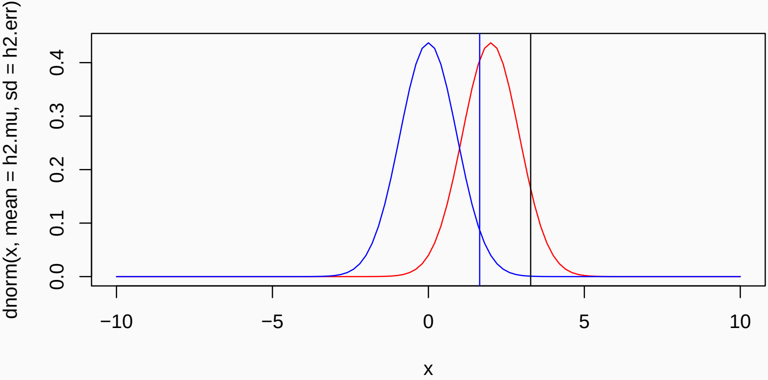
Valor Z calculado:

```
z<- ((mean(x2)-mean(x1)) - 0)/ (5/sqrt(30)); z

## [1] 3.279533
```

Valor Z associado à probabilidade de 0.05:

```
z_005 <- qnorm(0.05,0,1,lower.tail=F); z_005

## [1] 1.644854
```

```
curve(dnorm(x, mean = h2.mu, sd = h2.err), -10, 10, , col = "red")
curve(dnorm(x, mean = h1.mu, sd = h1.err), -10, 10, , col = "blue", add = T)
abline(v = z_005, col = "blue")
abline(v = z, lty = 1)
```

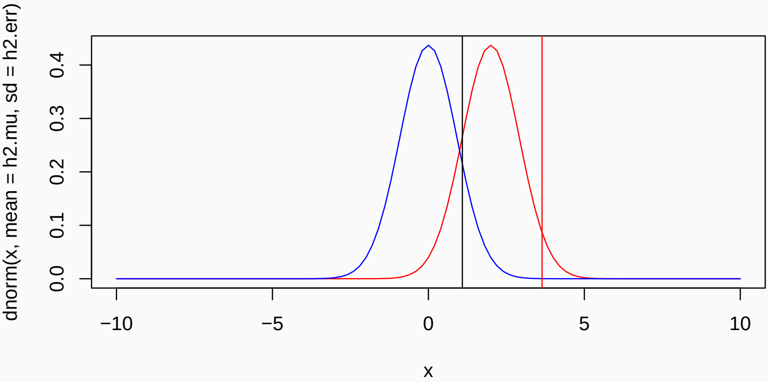


Apesar de não ser comum, nada impede que testemos diferentes hipóteses ”competitivas”, por exemplo, $H_2 : \mu_D = 2$

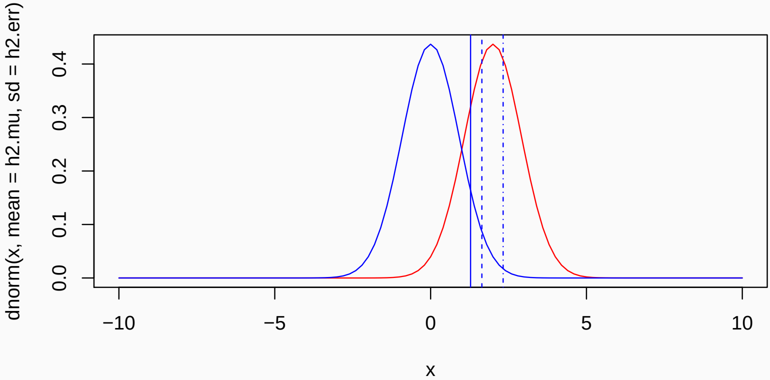
```
z<- ((mean(x2)-mean(x1)) - 2)/ (5/sqrt(30)); z

## [1] 1.088642

z_005 <- qnorm(0.05,2,1,lower.tail=F); z_005

## [1] 3.644854
```

ERROS TIPO I (α)

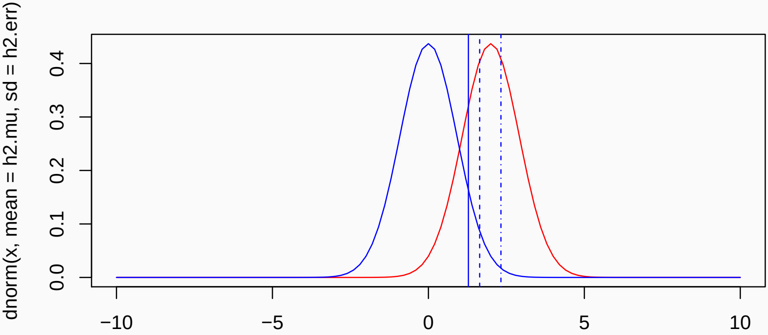
```
curve(dnorm(x, mean = h2.mu, sd = h2.err), -10, 10, , col = "red")
curve(dnorm(x, mean = h1.mu, sd = h1.err), -10, 10, , col = "blue", add = T)
abline(v = z_005, col = "red")
abline(v = z, lty = 1)
```



O erro Tipo I, ou erro α, é a probabilidade de rejeitarmos H_1 em favor de H_2, quando H_1 é verdadeira. Ao estabelecermos um nível de significancia, decidimos qual porcentagem de erro Tipo I é aceitável:



Contudo, podemos observar que existe um outro tipo possível de erro: rejeitar H_2 em favor de H_1, quando na verdade H_2 é verdadeira. Esse é o chamado erro Tipo II (β), também conhecido como **poder (*power*)** do teste.



ERROS TIPO II (β)

E agora? Se aumentamos o nível de significância, perdemos poder.

Seria esse o momento de abandonar de vez a ciência, e vender arte na praia?

ERROS TIPO II (β)

E agora? Se aumentamos o nível de significância, perdemos poder.

Seria esse o momento de abandonar de vez a ciência, e vender arte na praia?

Como poderíamos resolver esse problema?

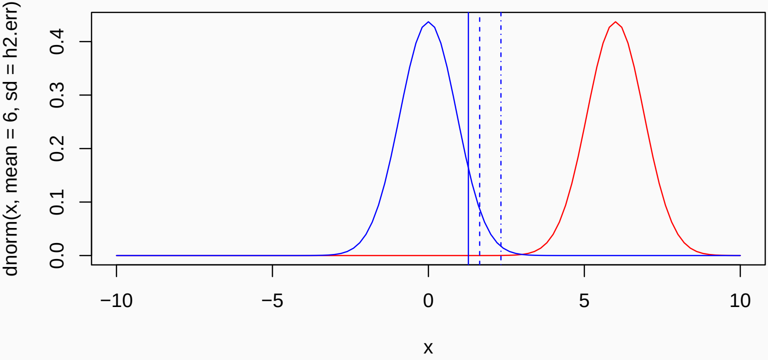
E agora? Se aumentamos o nível de significância, perdemos poder.

Seria esse o momento de abandonar de vez a ciência, e vender arte na praia?

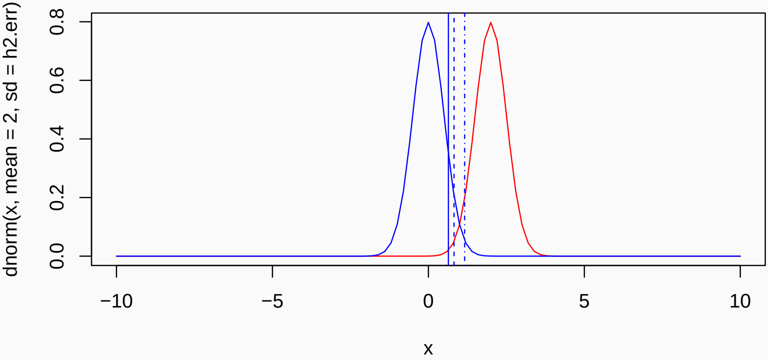
Como poderíamos resolver esse problema?

1. Avaliando efeitos maiores
2. Reduzindo a nossa variância

Avaliando efeitos maiores: $\mu_D = 6$



Aumentando a amostragem: n = 100, $\sigma_\mu = \dfrac{\sigma}{\sqrt{N}}$



Esta relação nos permite estimar o esforço amostral necessário para garantir que tenhamos poder estatístico suficiente para detectar um efeito de tamanho d, com base na definição dos erros α e β e em uma estimativa de σ.

A maneira ideal de determinar os parâmetros necessários é a realização de um estudo piloto. Mas podemos também recorrer à literatura e/ou ao bom senso.