Distribuições Amostrais

Distribuição Amostral da Média

- Teorema Central do Limite

Prof. Aloisio Ribeiro
Departamento de Estatística - UFMG

Teorema Central do Limite

Na última aula, vimos que, para amostras aleatórias de tamanho de tamanho n obtidas de uma população com média μ e variância σ^2, a distribuição de $\bar{X}$ simétrica possui

$$E(\bar{X}) = \mu \qquad \text{e} \qquad VAR(\bar{X}) = \frac{\sigma^2}{n}$$

O **Teorema Central do Limite (T.C.L)** nos assegura que, para grandes amostras, a distribuição de $\bar{X}$ pode ser aproximada por uma distribuição Normal, qualquer que seja a distribuição da variável X.

Teorema Central do Limite

Seja uma amostra aleatória $X_1, X_2, \ldots, X_n$, de uma variável aleatória X com média μ e desvio padrão σ. À medida que n cresce, a distribuição de probabilidade da média amostral $\bar{X}$ aproxima-se de uma Normal com média μ e desvio padrão $\frac{\sigma}{\sqrt{n}}$.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$ aproxima-se da distribuição N(0,1)

Corolário: para amostras aleatórias de uma população Normal, a distribuição amostral da média é **exatamente Normal** para qualquer tamanho de amostra n

Importante: Em geral, a aproximação da distribuição da média amostral pela distribuição $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ é satisfatória quando o tamanho de amostra n é maior ou igual 30.

Como simular valores para uma variável aleatória X, com certa distribuição de probabilidade?

- A ordem do quantil é uma variável aleatória uniforme no intervalo (0,1)

- Gere aleatoriamente n valores entre 0 e 1 (ordens dos quantis)

- Para cada valor gerado da ordem do quantil, obtenha o quantil correspondente da distribuição de interesse (valore gerado para X

Exemplo: Gere 10 valores de uma N(50 , 5) usando o R

```
# gera 10 valores para a ordem dos quantis

ordem = runif(10, 0,1)

# Obtenha os quantis da distribuição normal para cada ordem de quantil

x  = qnorm(ordem, mean=50, sd=5)
```

```
> ordem
 [1] 0.74424179 0.21878950 0.43579423 0.91961888 0.36711288
 [6] 0.34480013 0.05445047 0.78973118 0.32323968 0.43281579

> x
 [1] 53.28239 46.11856 49.19179 57.01256 48.30245 48.00301
 [7] 41.98424 54.02744 47.70671 49.15395
```

- Os valores de X podem ser gerados com um único comando

```
x = rnorm(10, mean=50, sd=5)
```

```
> x
 [1] 41.77634 55.51846 45.80427 41.77516 61.53235 52.96888
 [7] 37.31797 47.96718 49.11799 50.46371
```

- A letra r em rnorm é de Random (aleatório)
- Para cada distribuição de probabilidade, existe no R uma função similar

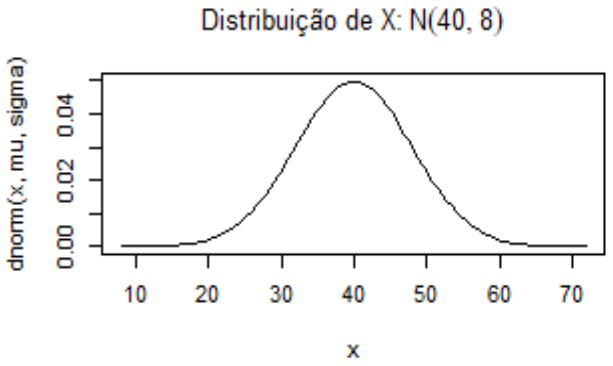
  rbinom para gera valores de uma distribuição B(n,p)

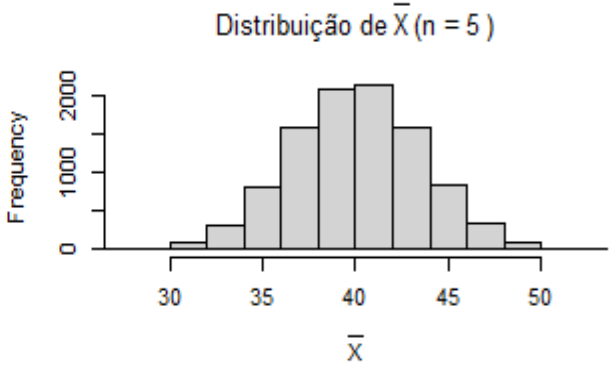
  rpois para gerar valore de uma distribuição poisson

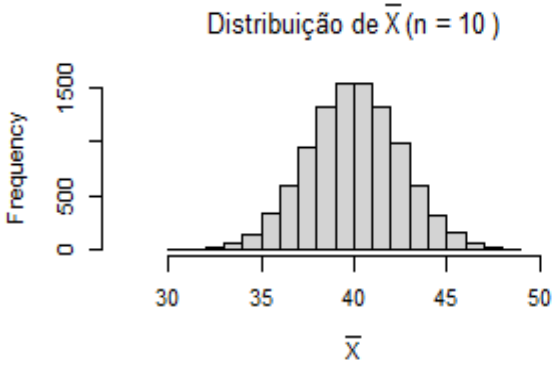
Ilustração do Teorema Central do Limite

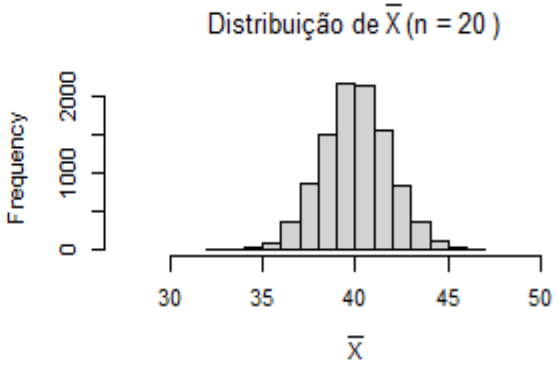
Amostras de uma população N(40,8)

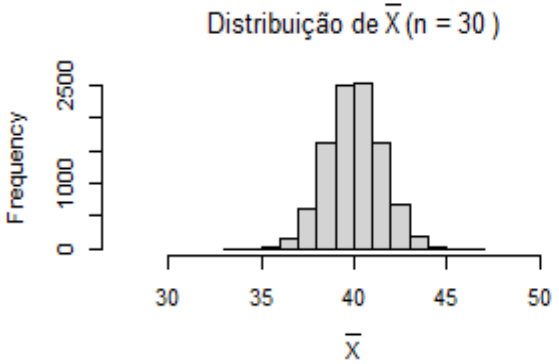
1) Obtenha por simulação uma amostra de 5 valores de uma distribuição N(40,8).
2) Calcule a média desses 5 valores e guarde seu valor
3) Repita os passos 1 e 2 1000 vezes, obtendo desta forma, por simulação, 1000 valores da média amostral para amostras de tamanho 5.
4) Calcule a média e o desvio padrão para estes 1000 valores simulados da média amostral. Compare esses valores com os valores teóricos.
5) Com esses 1000 valores construa o histograma e verifique se ele se assemelha ao histograma de uma distribuição Normal.
6) Repita os passos de 1 a 5 para valores de n = 10, 20, 30 e 50.
7) Observe o que acontece com o histograma quando n cresce.











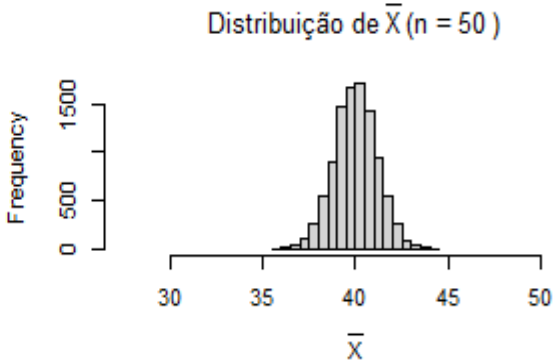


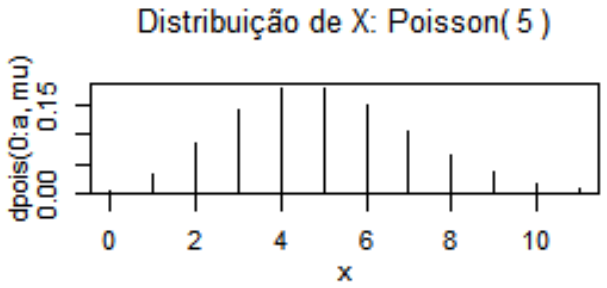
Ilustração do T.C.L
Caso de uma população N(40, 8)

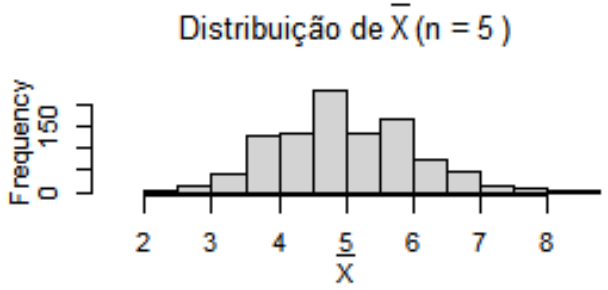
n	Valores teóricos		Valores estimados por simulação	
	$E(\bar{X})$	$DP(\bar{X})$	$E(\bar{X})$	$DP(\bar{X})$
5	40	3,58	40,04	3,54
10	40	2,53	40,01	2,53
20	40	1,79	39,99	1,80
30	40	1,46	40,01	1,46
50	40	1,13	39,99	1,14

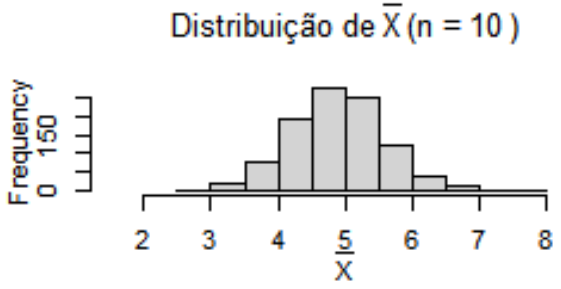
```
# simula uma amostra de tamanho n de uma N(um,sigma) e calcula sua média
simula = function(n, mu, sigma){
x = rnorm(n, mu, sigma)
media = mean(x)}

# simula ns amostras de tamanho n de uma N(um, sigma)
TCLNormal=function(nsimul, n, mu, sigma){
medias = replicate(n=nsimul, simula(n,mu,sigma))
resultado=c(n, mean(medias), var(medias), sd(medias))
names(resultado)=c("n", "média", "variância", "desvio padrão")
hist(medias, xlim=c(mu-3.5*sigma/sqrt(5),mu+3.5*sigma/sqrt(5)),  xlab=expression(bar(X)),main=NULL)
title(bquote("Distribuição de" ~bar(X)~ "(n =" ~.(n)~ ")"))
return(resultado)}

# obtendo os resultados para diferentes valores de n
par(mfrow=c(3,2))
mu = 40
sigma = 8
curve(dnorm(x,mu,sigma),mu-4*sigma,mu+4*sigma, main = bquote("Distribuição de X:"~N(.(mu),.(sigma))))
TCLNormal(ns = 1000, n = 5, mu = 40, sigma = 8 )
TCLNormal(ns = 1000, n = 10, mu = 40, sigma = 8 )
TCLNormal(ns = 1000, n = 20, mu = 40, sigma = 8 )
TCLNormal(ns = 1000, n = 30, mu = 40, sigma = 8 )
TCLNormal(ns = 1000, n = 50, mu = 40, sigma = 8 )
```

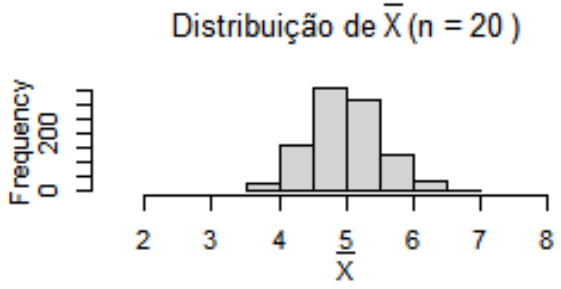
Execute esse código no R alterando os valores de mu e sigma para mu = 100 e sigma = 20.

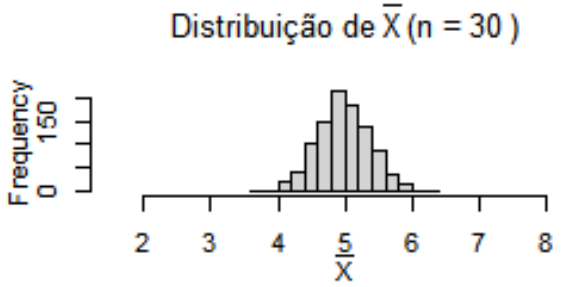
Ilustração do T.C.L.
Caso de uma população Poisson(5)

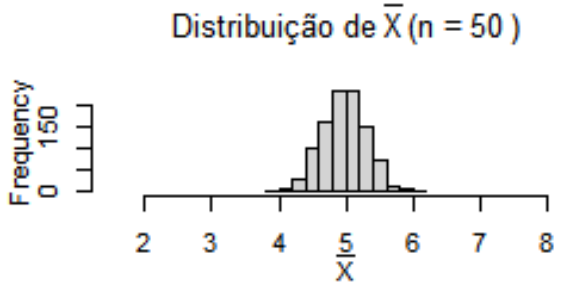








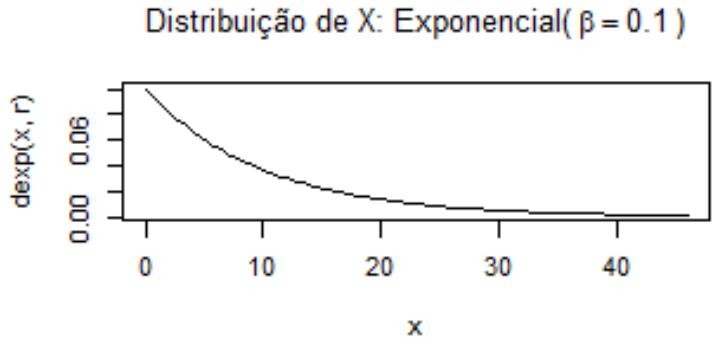


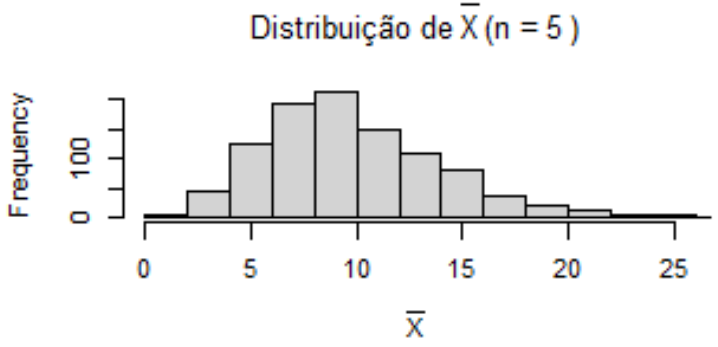


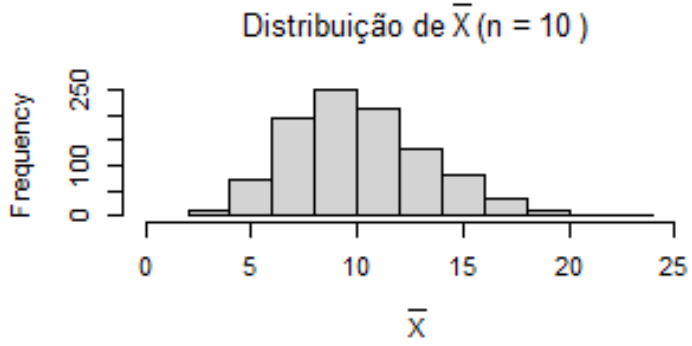
n	Valores teóricos		Valores obtidos por simulação	
	$E(\bar{X})$	$DP(\bar{X})$	$E(\bar{X})$	$DP(\bar{X})$
5	5	1	5,03	1,01
10	5	0,71	5,00	0,70
20	5	0,50	4,98	0,49
30	5	0,41	5,02	0,41
50	5	0,32	4,99	0,31

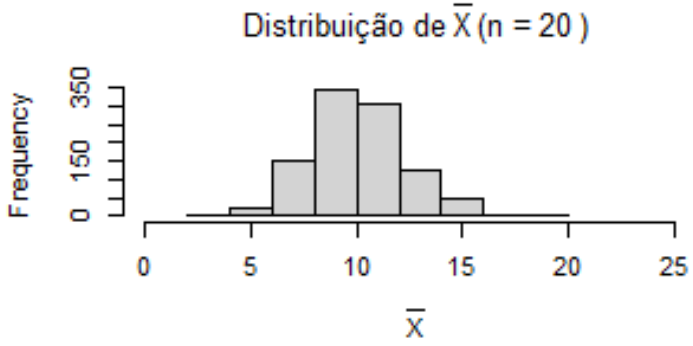
```
# simula uma amostra de tamanho n de uma Poisson(mu) e calcula sua média
simula = function(n, mu){
x = rpois(n, mu)
media = mean(x)}
# simula ns amostras de tamanho n de uma Poisson(mu)
TCLPois=function(nsimul, n, mu){
medias = replicate(n=nsimul, simula(n,mu))
resultado=c(n, mean(medias), var(medias), sd(medias))
names(resultado)=c("n", "média", "variância", "desvio padrão")
hist(medias, xlim=c(mu-3.5*sqrt(mu/5),mu+3.5*sqrt(mu/5)),  xlab=expression(bar(X)),main = NULL)
title(bquote("Distribuição de" ~ bar(X) ~ "(n =" ~.(n)~ ")"))
return(resultado)}
# obtendo os resultados para diferentes valores de n
par(mfrow=c(3,2))
mu = 5
a = qpois(0.99, mu)
plot(0:a, dpois(0:a,mu),type="h", xlab="x")
title(main = bquote("Distribuição de X: Poisson(" ~ .(mu) ~ ")"))
TCLPois(ns = 1000, n = 5, mu)
TCLPois(1000, 10, mu)
TCLPois(1000, 20, mu)
TCLPois(1000, 30, mu)
TCLPois(1000, 50, mu)
```

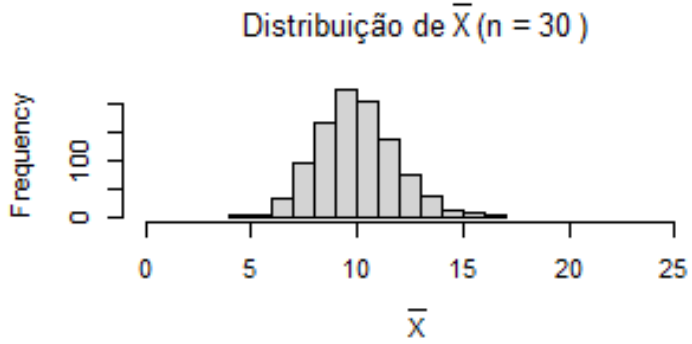
Execute esse código no R alterando o valor de mu = 100.











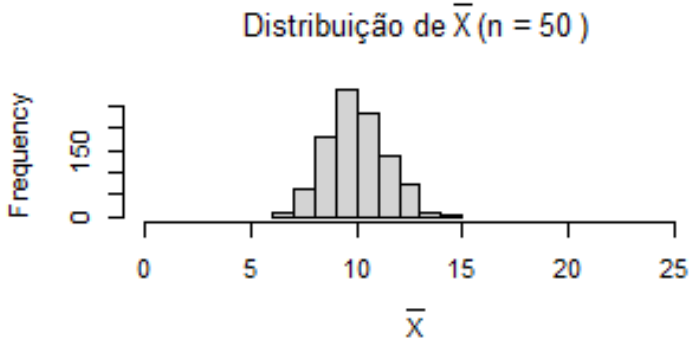


Ilustração do T.C.L
Caso de uma população Exponencial
Com média igual a 10

n	Valores teóricos		Valores obtidos por simulação	
	$E(\bar{X})$	$DP(\bar{X})$	$E(\bar{X})$	$DP(\bar{X})$
5	10	4,47	9,89	4,58
10	10	3,16	10,11	3,14
20	10	2,24	9,94	2,20
30	10	1,83	9,95	1,82
50	10	1,41	10,02	1,42

Neste caso, para n = 20, a distribuição da média amostral já é bem próxima de uma distribuição Normal.

```
# simula uma amostra de tamanho n de uma Exponencial com média mu e calcula sua média
simula = function(n, mu){
r = 1/mu
x = rexp(n, r)
media = mean(x)}
# simula ns amostras de tamanho n de uma Poisson(mu)
TCLExp=function(nsimul, n, mu){
medias = replicate(n=nsimul, simula(n,mu))
resultado=c(n, mean(medias), var(medias), sd(medias))
names(resultado)=c("n", "média", "variância", "desvio padrão")
hist(medias, xlim=c(max(0,mu-3.5*mu/sqrt(5)),mu + 3.5*mu/sqrt(5)),  xlab=expression(bar(X)),main = NULL)
title(bquote("Distribuição de" ~ bar(X) ~ "(n =" ~.(n)~ ")"))
return(resultado)}
# obtendo os resultados para diferentes valores de n
par(mfrow=c(3,2))
mu = 10
r = 1/mu
curve(dexp(x, r), from=0, to=qexp(0.99,r))
title(main = bquote("Distribuição de X: Exponencial(" ~ beta == .(1/mu) ~ ")"))
TCLExp(ns = 1000, n = 5, mu)
TCLExp(1000, 10, mu)
TCLExp(1000, 20, mu)
TCLExp(1000, 30, mu)
TCLExp(1000, 50, mu)
```

Execute esse código no R alterando o valor de mu e sigma para mu = 20.

Exemplo: Suponha que o tempo de vida de um dispositivo possua uma distribuição de probabilidade com média e desvio padrão iguais a 50 horas. Se uma amostra aleatória de 40 dispositivos é observada, qual a probabilidade do tempo médio de duração dos dispositivos da amostra ser menor ou igual a 60 horas?

X – tempo de vida do dispositivo

E(X) = DP(X) = 50

Como n = 40 ≥ 30 podemos usar o T.C.L.

A distribuição de $\bar{X}$ é aproximadamente $N\left(50, \frac{50}{\sqrt{40}}\right)$

$$P(\overline{X} \leq 60) \cong P\left(Z \leq \frac{60-50}{50/\sqrt{40}}\right) = P(Z \leq 1{,}2649) = 0{,}8970$$

No R

pnorm(60, mean = 50, sd=50/sqrt(40))

$\bar{X} - \mu$ é chamado de erro amostral

mede o quanto estamos errado ao estimar a média populacional pela média amostral

$$Se\ \bar{X} \text{ é aproximadamente } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right),$$

$$então\ \ \bar{X} - \mu \ \text{ é aproximadamente } N\left(0, \frac{\sigma}{\sqrt{n}}\right)$$

Exemplo: Suponha que o tempo médio de vida dos dispositivos é desconhecido.
a) Ao estimá-lo através de uma amostra aleatória simples de tamanho n = 40, e assumindo $\sigma = 50,$ qual a probabilidade do erro amostral ser, em valor absoluto, no máximo igual a 3?

$$\bar{X} - \mu \ \text{ é aproximadamente } N\left(0, \frac{50}{\sqrt{40}}\right)$$

$$P(|\bar{X} - \mu| \leq 3) = P(-3 \leq \bar{X} - \mu \leq 3) \cong P\left(\frac{-3-0}{50/\sqrt{40}} \leq Z \leq \frac{3-0}{50/\sqrt{40}}\right) = P(-0{,}3794 \leq Z \leq 0{,}3794)$$

$$= P(Z \leq 0{,}3794) - P(Z \leq -0{,}3794) = 0{,}2956$$

Exemplo: Suponha que o tempo médio de vida dos dispositivos é desconhecido.

a) Ao estimá-lo através de uma amostra aleatória simples de tamanho n = 40, e assumindo $\sigma = 50,$ qual a probabilidade do erro amostral ser, em valor absoluto, no máximo igual a 3?

$$\bar{X} - \mu \quad \text{é aproximadamente } N\left(0, \frac{50}{\sqrt{40}}\right)$$

$$\text{P}(|\bar{X} - \mu| \leq 3) = \text{P}(-3 \leq \bar{X} - \mu \leq 3) \cong \text{P}\left(\frac{-3-0}{50/\sqrt{40}} \leq Z \leq \frac{3-0}{50/\sqrt{40}}\right) = \text{P}(-0{,}3794 \leq Z \leq 0{,}3794)$$

$$= \text{P}(Z \leq 0{,}3794) - \text{P}(Z \leq -0{,}3794) = 0{,}2956$$

b) Com tamanho de amostra igual a 40, e desvio padrão σ = 50, encontre o valor E que deixa 95% dos valores absolutos do erro amostral abaixo dele.

$$\text{Qual o valor E tal que } P(-E \leq \overline{X} - \mu \leq E) = 0{,}95$$

$$P(\overline{X} - \mu \leq E) = 0{,}975 \Rightarrow E = Q_{0{,}975}^{N\left(0,\, 50/\sqrt{40}\right)}$$

No R:
> qnorm(p = 0.975, mean = 0, sd = 50/sqrt(40))
[1] 15,49

- Com n = 40, o erro de estimação é, em valor absoluto, e com probabilidade igual a 0,95, no máximo igual a 15,49 horas.
- O erro máximo de estimação E é chamado de **margem de erro.**

Expressando E em função do quantil de ordem 0,975 da distribuição N(0,1)

$$E = Q_{0{,}975}^{N(0,1)} \frac{50}{\sqrt{40}} = 1{,}96\frac{50}{\sqrt{40}}$$

Para tamanho de amostra n, desvio padrão σ e probabilidade 1 - α

$$E = z_{\alpha/2}\ \frac{\sigma}{\sqrt{n}}$$

$Z_{\alpha/2}$ é o quantil de ordem $1 - \alpha/2$ da distribuição N(0,1)

Como reduzir o erro máximo de estimação?

Aumentando o tamanho da amostra

Exemplo: Qual o valor de n para que que com probabilidade 0,95 o erro amostral de estimação seja no máximo igual a 5 horas? Considere σ = 50.

Encontrar n tal que $\mathrm{E} = \ 5 = 1{,}96\ \frac{50}{\sqrt{n}}$ $\qquad n = \left(\frac{1{,}96\ \times 50}{5}\right)^2 = 384{,}16 \Rightarrow n \geq 385$

Caso geral: Qual o valor de n tal que com probabilidade 1 - α o erro amostral de estimação seja de no máximo E?

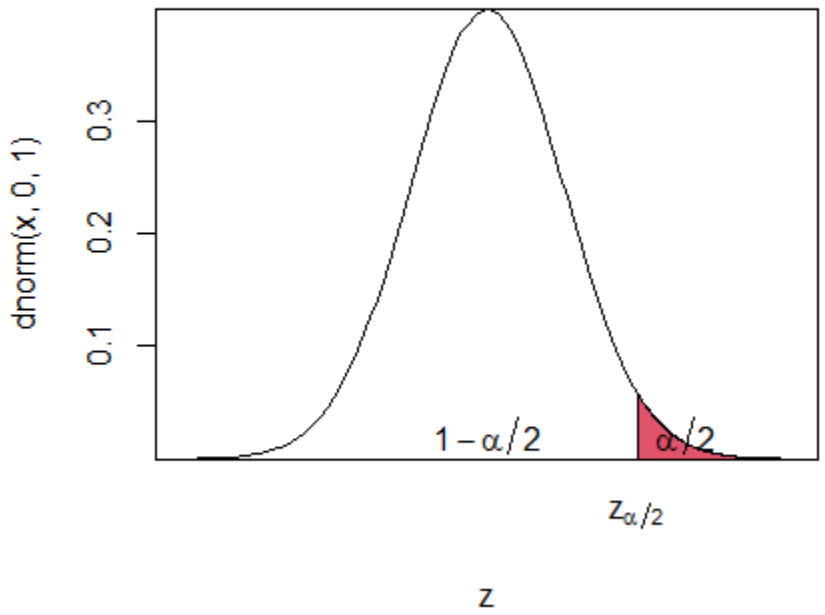
$$n \geq \left(\frac{z\alpha_{/2} \times \sigma}{E} \right)^2$$

$z_{\alpha/2}$ é o quantil de ordem $1 - \alpha/2$ *da distribuição* $N(0,1)$

- Quanto maior for a incerteza sobre a variável de interesse (σ) maior será o tamanho da amostra.
- Quanto maior for a precisão desejada (menor for o valor de E), maior será o tamanho da amostra.
- α e E são fixados pelo pesquisador.
- σ é geralmente desconhecido. Como obter o valor de σ?

Usando o conhecimento prévio da variável ou através de uma amostra piloto.

$z_{\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da distribuição $\mathrm{N}(0,1)$



$1-\alpha = 0{,}95$	$\alpha/2$ = 0,025	1 - $\alpha/2$ = 0,975	$z_{0{,}025} = 1{,}96$	No R: qnorm(0.975)
$1-\alpha = 0{,}90$	$\alpha/2$ = 0,05	1 - $\alpha/2$ = 0,95	$z_{0{,}05} = 1{,}64$	No R: qnorm(0.95)
$1-\alpha = 0{,}99$	$\alpha/2$ = 0,005	1 - $\alpha/2$ = 0,995	$z_{0{,}005} = 2{,}58$	No R: qnorm(0.995)

Formula alternativa para n

- Fixar E como uma fração do desvio padrão **σ:** **E = k σ**

$$n \geq \left(\frac{z\alpha/_2 \times \sigma}{k\sigma}\right)^2 = \left(\frac{z\alpha/_2}{k}\right)^2$$

Exemplo: E = 0,15 σ $\quad$ $n \geq \left(\frac{1,96}{0,15}\right)^2 = 170,74$

As fórmulas apresentadas para n valem para populações infinitas, onde não se pode mensurar o tamanho da população, ou para situações onde o tamanho da amostra é muito pequeno relativo ao tamanho da população. Para populações finitas, existem fatores de correção que devem ser aplicados aos tamanhos de amostras obtidos com essas fórmulas. Estes fatores são função do tamanho da população, N.