Estatística e Informática

Aula 09 - Estatística, Distribuição Amostral

Alan Rodrigo Panosso alan.panosso@unesp.br

Departamento e Ciências Exatas FCAV/UNESP

(02-05-2024)

Distribuições Amostral

Parâmetro e Estatística

Parâmetro: é uma medida usada para descrever uma característica da população.

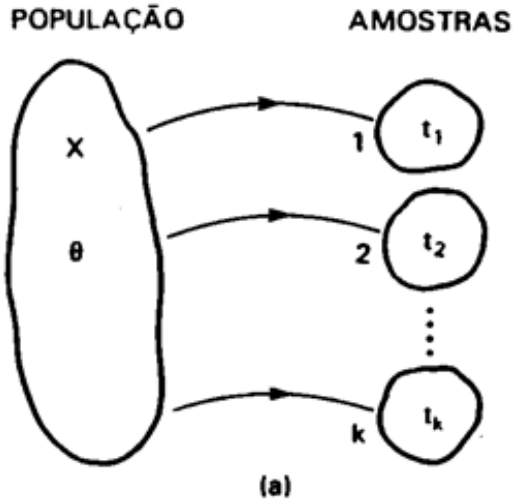
Estatística ou **Estimador**: é qualquer função de uma amostra aleatória (fórmula ou expressão), construída com o propósito de servir como instrumento para descrever alguma característica da amostra e para fazer *inferência* a respeito da característica na população.

Resumo	Parâmetro	Estatística
Média	μ	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variância	σ^2	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$
Proporção	π	$\hat{p} = \frac{X}{n}$
Correlação	ρ	$r = \frac{\sum_{i=1}^{n}(x_1-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_1-\bar{x})^2(y_i-\bar{y})^2}}$

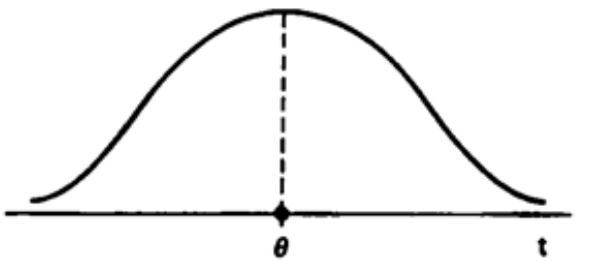
Toda **estatística**, sendo uma função de uma amostra aleatória $X_1, X_2, \cdots X_n$, é também uma variável aleatória e tem uma distribuição.

Assim, o comportamento da estatística pode ser descrito por alguma distribuição de probabilidade, agora denominada **distribuição amostral**.

Assim, cada **estatística** é uma variável aleatória e sua distribuição de probabilidade é chamada distribuição amostral da estatística.



(a)



(b)

Distribuições Amostral da Média

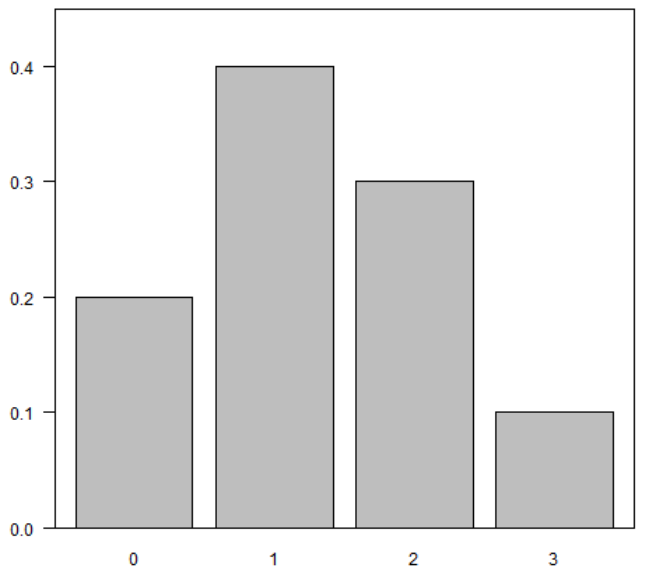
Seja a variável aleatória X que denota o tempo em dias de germinação de sementes de uma espécie vegetal, após a semeadura.

Considerando a população de todas sementes dessa espécie, suponha que X tem a distribuição de probabilidade apresentada abaixo.

Uma amostra aleatória simples **COM REPOSIÇÃO** (X_1, X_2) de 2 covas $(n = 2)$ é tomada nesta área. Qual a distribuição do tempo médio amostral em dias para a germinação?

x	0	1	2	3
$P(X)$	$0,20$	$0,40$	$0,30$	$0,10$

Distribuição de Probabilidade



Calculando a Esperança de X:

$$E(X) = \sum_{i=1}^{k} x_i \cdot P(x_i) = 1,30\ dias$$

Calculando a Variância de X.

$$Var(X) = \sum_{i=1}^{k} x_i^2 \cdot P(x_i) - [E(X)]^2$$

$$Var(X) = 0,81\ dias^2$$

Definimos a média a partir da tomada de duas amostras temos:

$$\bar{X} = \frac{X_1 + X_2}{2}$$

De acordo com a definição de amostra aleatória simples com reposição, X_1 e X_2 são variáveis aleatórias independentes, assim, a distribuição conjunta pode ser obtida multiplicando-se as probabilidades marginais.

$$P(0 \cap 1) = P(0) \times P(1) = 0,2 \times 0,4 = 0,08$$

Tabela com os possíveis valores de médias $\bar{X}$, dados os valores de X.

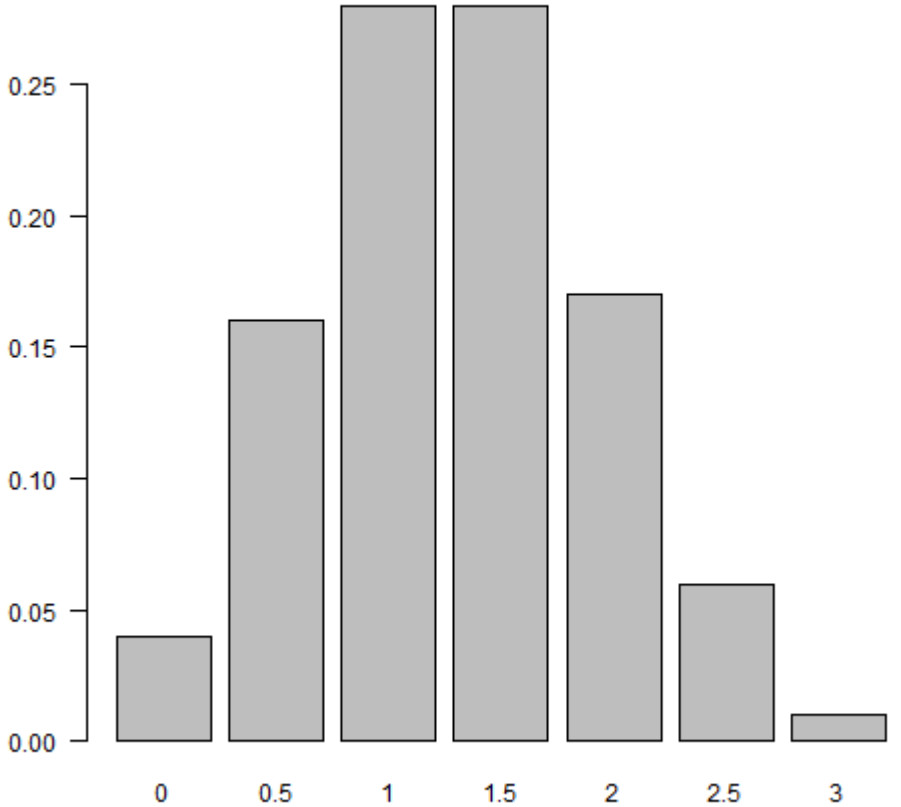
X_1/X_2	**0**	**1**	**2**	**3**
0	0,00	0,50	1,00	1,50
1	0,50	1,00	1,50	2,00
2	1,00	1,50	2,00	2,50
3	1,50	2,00	2,50	3,00

Tabela com a distribuição conjunta de X_1 e X_2, ou seja, as probabilidades $P(\bar{X})$.

X_1/X_2	**0**	**1**	**2**	**3**	$\sum$
0	0,04	0,08	0,06	0,02	**0,2**
1	0,08	0,16	0,12	0,04	**0,4**
2	0,06	0,12	0,09	0,03	**0,3**
3	0,02	0,04	0,03	0,01	**0,1**
$\sum$	**0,2**	**0,4**	**0,3**	**0,1**	**1,00**

Distribuição de probabilidade de $\bar{X}$.

$\bar{X}$	0	$0,5$	1	$1,5$	$2,0$	$2,5$	$3,0$
$P(\bar{X})$	$0,04$	$0,16$	$0,28$	$0,28$	$0,17$	$0,06$	$0,01$



Cálculo da Esperança e da Variância da Média do Número de dias para Germinação de X $(n = 2)$.

$$E(\bar{X}) = \sum_{n=1}^{k} = \bar{x_i} \cdot P(\bar{x_i}) = 1,30\ dias$$

$$Var(\bar{X}) = E(\bar{X}^2) - [E(\bar{X})]^2$$

$$Var(\bar{X}) = 2,095 - (1,3)^2$$

$$Var(\bar{X}) = 0,405\ dias^2 = \frac{\sigma^2}{n} = \frac{0,81}{2} = \frac{Var(X)}{n}$$

Se $X_1, X_2, \cdots, X_n$ constitui uma amostra aleatória simples com reposição de uma população que tem média μ e variância σ^2, então:

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

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

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

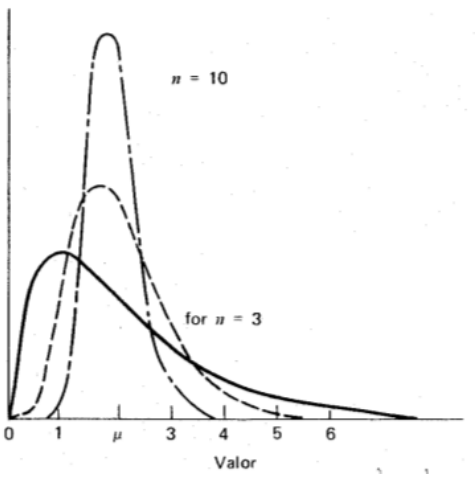
Observações importantes

- O desvio padrão da média $DP(\bar{X})$ e o erro padrão da média $s(m)$ são termos equivalentes.
- O erro padrão da média é geralmente usado para evitar confusão com o desvio padrão (σ) das observações.
- O aumento do tamanho da amostra (n) o desvio padrão da distribuição da média diminui, portanto os valores de $\bar{X}$ são mais próximos à μ.
- Estima-se o erro padrão da média usando o tamanho da amostra (n) e desvio padrão (s) de uma única amostra de observações.
- À medida que o tamanho da amostra aumenta temos uma estimativa mais precisa do desvio padrão paramétrico (σ) da população.
- Em contraste, o erro padrão da média torna-se uma estimativa mais precisa da média paramétrica (μ), pois o erro padrão da média diminui.

Teorema Central do Limite

Se $\bar{X}$ é a média de uma amostra aleatória simples com reposição, de tamanho n, de uma população **normal**, com média μ e variância σ^2, sua distribuição é **normal**, com média μ e variância $\frac{\sigma^2}{n}$.

Assim, em uma amostra aleatória simples com reposição, de tamanho n, de uma população arbitrária, com média μ e variância σ^2, a distribuição de $\bar{X}$ quando n é grande é **aproximadamente normal**, com média μ e variância $\frac{\sigma^2}{n}$.



Indução do conceito a partir do R

Simulação de ume população não normal.

```
set.seed(7355608)
x <- rep(1:10,
                 c(800,700,200,350,120,
                   500,1200,800,800,800)
                 )
```

Cálculo da esperança de X: $E(X) = \mu$**.**

```
mean(x)
```

```
#> [1] 6.028708
```

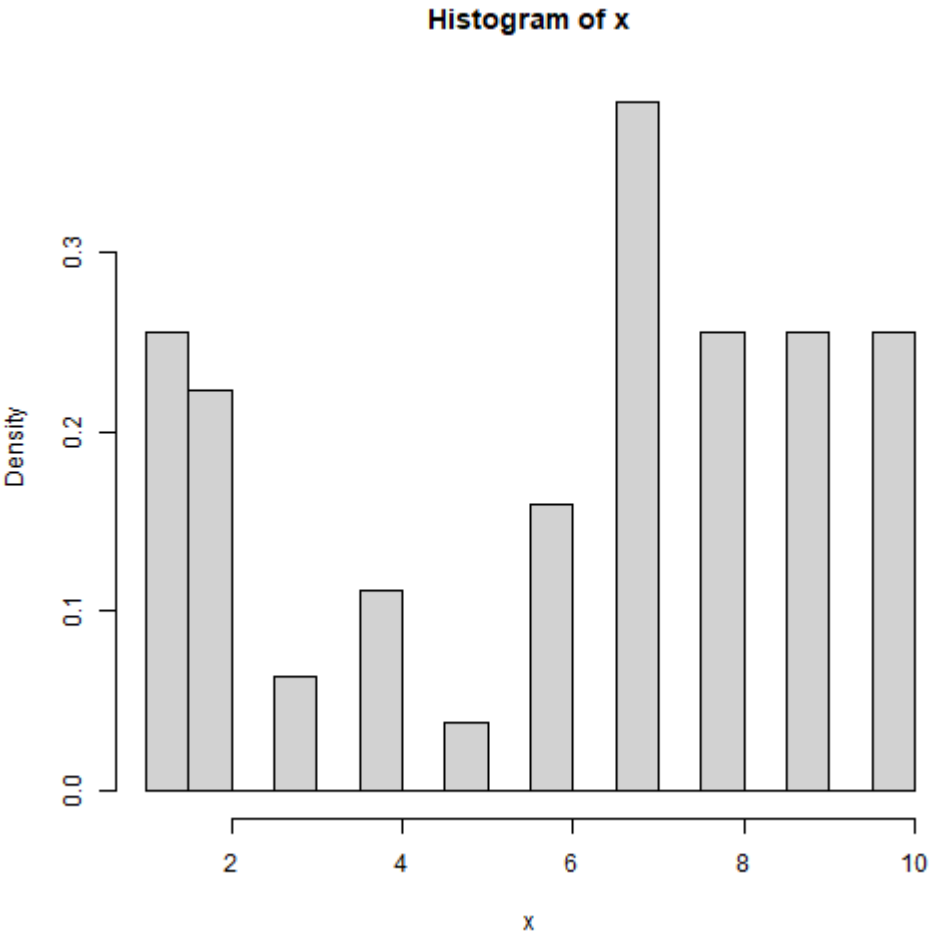
Cálculo da variância de X: $Var(X) = \sigma^2$**.**

```
var(x)
```

```
#> [1] 9.397804
```

Distribuição de probabilidade.

```
hist(x, probability = TRUE)
```



Definindo o tamanho da amostra n.

```
n <- 5
```

Realizar uma amostragem aleatória simples, com reposição, de tamanho n e calcular a média e variância dessa amostra.

```
amostra <- sample(x,n,replace = TRUE)
mean(amostra)
```

```
#> [1] 5.4
```

```
var(amostra)
```

```
#> [1] 9.3
```

Realizar k amostragens de tamanho n, calcular as respectivas médias, construir a distribuição de probabilidade dessas médias com os cálculos da $E[\bar{X}]$ e da $Var(\bar{X})$.

```
k <- 10000
vetor_medias <- 0
for(i in 1:k) vetor_medias[i] <- mean(sample(x,n,replace=TRUE))
```

Cálculo da esperança de X: $E(\bar{X}) = \mu$.

```
mean(vetor_medias)
```

```
#> [1] 6.02868
```

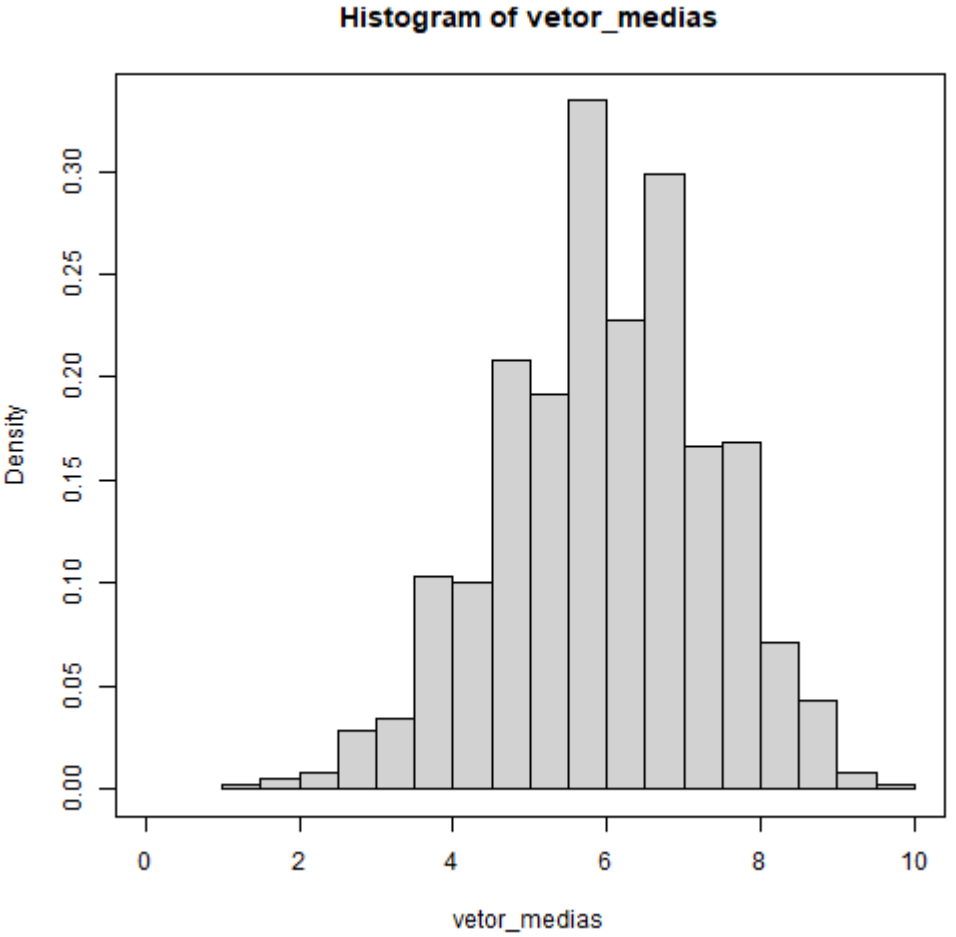
Cálculo da variância de X: $Var(\bar{X}) = \frac{\sigma^2}{n}$.

```
var(vetor_medias)
```

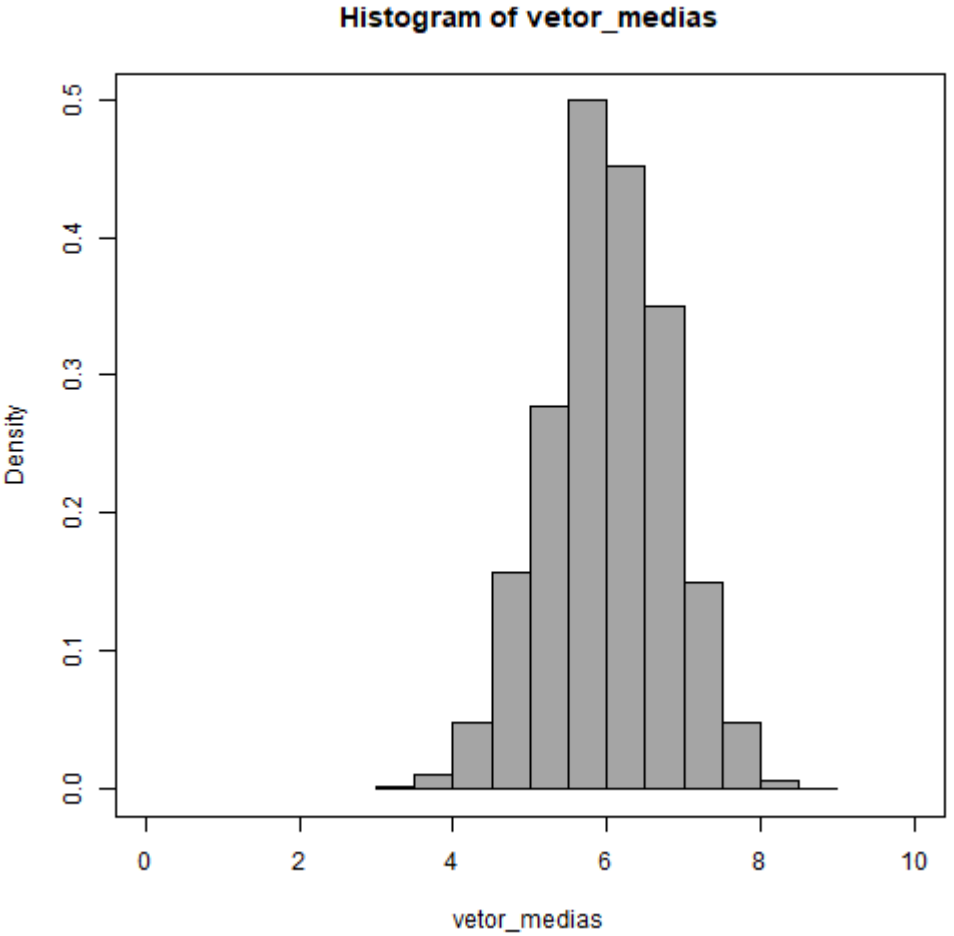
```
#> [1] 1.902184
```

```
hist(vetor_medias,xlim=c(0,10), probability = TRUE);box()
```

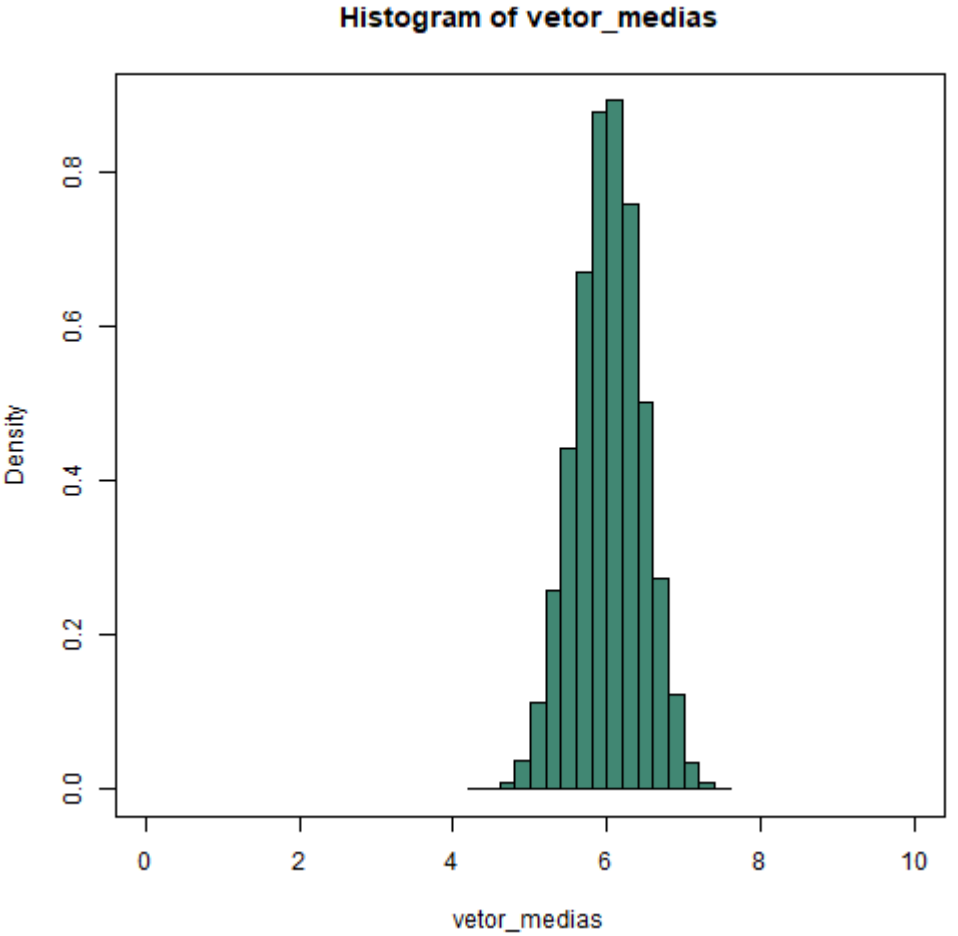
n = 5


Histogram of vetor_medias
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
2
4
6
8
10
vetor_medias

n = 15


Histogram of vetor_medias
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
vetor_medias

n = 50


Histogram of vetor_medias
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
vetor_medias

Acesse o Link para simularmos essa amostragem. Aba "Distribuição Amostral - Média"

https://arpanosso.shinyapps.io/estatinfo/

Transformação normal padrão para cálculo das probabilidades

Assim, podemos utilizar a tabela normal padrão para o cálculo das probabilidades associadas aos intervalos de $\bar{X}$ a partir da transformação:

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

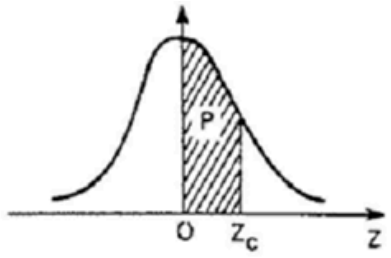
onde $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ e n é o tamanho da amostras aleatória simples tomada da população.

Tabela - Normal Padrão

TÁBUA III

Distribuição normal reduzida: N(0;1)

Probabilidades p tais que $p = P(0 < Z < Z_c)$



parte inteira e primeira decimal de Z_c	SEGUNDA DECIMAL DE Z_c										parte inteira e primeira decimal de Z_c
	0	1	2	3	4	5	6	7	8	9	
	p = 0										
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586	0,0
0,1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535	0,1
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409	0,2
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173	0,3
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793	0,4
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240	0,5
0,6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490	0,6
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524	0,7
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327	0,8
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891	0,9
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214	1,0
1,1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298	1,1
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147	1,2
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774	1,3
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189	1,4
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408	1,5
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449	1,6
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327	1,7

Exercício

Seja uma máquina de empacotar soja, cujos pesos das sacas (em kg) seguem uma distribuição $N(50, 2)$. Assim, se a máquina estiver regulada, qual a probabilidade de, colhendo-se uma amostra ao acaso de 100 sacas, observarmos uma média diferente de $50\ kg$ em menos de $0,2828\ kg$?

Resposta:

$$P(50 - 0,2828 \leq \bar{X} \leq 50 + 0,2828) = 0,9545$$

Amostras sem reposição de populações finitas

Supondo uma população com N elementos, se a amostragem for feita **SEM REPOSIÇÃO**, temos as esperança e a variância definidas como:

Esperança

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

Variância

$$Var(\bar{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Observações,

$\mu = E(X)$, $\sigma^2 = Var(X)$ e $\frac{N-n}{N-1}$ é o fator de correção para populações finitas.

A variância da média amostral com este tipo de amostragem é menor do que com reposição. Assim, amostragem sem reposição é mais eficiente do que a com reposição para estimar o valor médio (μ). No entanto, se a população for grande quando comparada com o tamanho da amostra (n), o fator de correção será próximo de 1 e, portanto:

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

Distribuições Amostral da Proporção

Distribuição Amostral da Proporção

Anteriormente, definimos X uma $v.a.$ para cada ensaio de **Bernoulli**, com $P(S) = p$.

Neste contexto, considerando n ensaios independentes, $X_1, X_2, \ldots, X_n$ constitui uma amostra *aleatória simples com reposição*. Como os resultados individuais são 0 (fracasso) ou 1 (sucesso), o número **TOTAL** de sucessos em n ensaios, é dado por:

$$T = \sum_{i=1}^{n} X_i$$

Portanto, a proporção amostral de suscessos em n ensaios é dada por:

$$\bar{X} = \hat{p} = \frac{T}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Observe que T tem distribuição binomial com parâmetros n e p, cuja esperança (média) é $n.p$ e variância $n.p.q$ ou $n.p.(1-p)$.

Então, para a proporção, temos:

Esperança	**Variância**
$E(\hat{p}) = E\left(\frac{T}{n}\right)$	$Var(\hat{p}) = Var\left(\frac{T}{n}\right)$
$E(\hat{p}) = \frac{1}{n}E(T)$	$Var(\hat{p}) = \frac{1}{n^2}Var(T)$
$E(\hat{p}) = \frac{1}{n}n.p$	$Var(\hat{p}) = \frac{1}{n^2}n.p.q$
$E(\hat{p}) = p$	$Var(\hat{p}) = \frac{p.q}{n}$

Assim, pelo **Teorema Central Limite**, quando n é grande, a proporção amostral $\hat{p}$ de sucessos em n ensaios de **Bernoulli** tem distribuição aproximadamente normal com média p e variância $\frac{p.q}{n}$, assim, podemos definir a variável padronizada Z como:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}} \text{ ou } Z = \frac{T - n.p}{\sqrt{n.p.q}}$$

Com distribuição aproximadamente normal padrão $N(0, 1)$.

Exercício

Em um ensaio experimental 625 covas foram semeadas com sementes comerciais provenientes de um lote com índice de germinação (p) igual a 70%.

Qual a probabilidade de se encontrar mais de 72% das covas com plantas germinadas?

Resposta:

$P(T \geq 450) = P(\hat{p} \geq 0,72) = P(Z \geq 1,09) = 0,13786$