2. Distribuições amostrais

USP-ICMC-SME

2013

Amostra aleatória

Notação. X: variável aleatória (v.a.).
$f(x;\theta)$: função densidade de probabilidade (X contínua) ou função massa de probabilidade (X discreta). Será denominada função densidade. θ é o parâmetro (pode ser um vetor). Se X é discreta, $f(x;\theta) = \mathrm{P}(X = x;\theta)$.

Definição

As v.a. aleatórias $X_1, X_2, \ldots, X_n$ constituem uma amostra aleatória de tamanho n de uma população com função densidade $f(x;\theta)$, se
(a) as n variáveis são independentes e
(b) cada X_i tem distribuição com função densidade $f(x;\theta)$.

Também chamada de amostra aleatória simples (AAS).

A definição de amostra aleatória é satisfeita quando a a população é infinita ou quando a população é finita e a amostra é selecionada com reposição.

Exemplo

Seleção de uma amostra de tamanho $n = 8$ de uma v.a. com distribuição Poisson(3). Em R: `rpois(8, 3)`.

Amostras selecionadas sem reposição de uma população finita, não satisfazem a definição da amostra aleatória, pois as variaveis aleatórias $X_1, \ldots, X_n$ não são independentes.
Se o tamanho da amostra é muito pequeno em relação ao tamanho da população, a definição é satisfeita aproximadamente.

Definição

Uma estatística T é qualquer função que dependa apenas da amostra $X_1, X_2, \ldots, X_n$.
$T = T(X_1, X_2, \ldots, X_n)$ com valores $t = T(x_1, \ldots, x_n)$.

Exemplo

$$\textit{Total amostral} : \sum_{i=1}^{n} X_i.$$

$$\textit{Média amostral} : \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

$$\textit{Desvio padrão amostral} : S = \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \right\}^{1/2}.$$

$$\textit{Máximo amostral} : X_{(n)} = \max(X_1, X_2, \ldots, X_n).$$

$$\textit{Amplitude amostral} : X_{(n)} - X_{(1)}, \quad X_{(1)} = \min(X_1, X_2, \ldots, X_n).$$

Uma estatística T é uma variável aleatória (pois é uma função das v.a. $X_1, X_2, \ldots, X_n$).
A distribuição de T é chamada de distribuição amostral da estatística T.

Exemplo

Distribuição do número de erros de impressão.

x	0	1	2	Total
$f(x) = P(X = x)$	1/2	2/5	1/10	1

Determine a distribuição do número médio de erros em amostras de $n = 2$ observações.

$\overline{X} = (X_1 + X_2)/2$.

Distribuição de (X_1, X_2) e cálculo de $\overline{x}$.

x_1, x_2	$\overline{x}$	$\mathsf{P}(X_1 = x_1, X_2 = x_2)$
0, 0	0	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
0, 1	1/2	$\frac{1}{2} \times \frac{2}{5} = \frac{1}{5}$
0, 2	1	$\frac{1}{2} \times \frac{1}{10} = \frac{1}{20}$
1, 0	1/2	$\frac{2}{5} \times \frac{1}{2} = \frac{1}{5}$
1, 1	1	$\frac{2}{5} \times \frac{2}{5} = \frac{4}{25}$
1, 2	3/2	$\frac{2}{5} \times \frac{1}{10} = \frac{1}{25}$
2, 0	1	$\frac{1}{10} \times \frac{1}{2} = \frac{1}{20}$
2, 1	3/2	$\frac{1}{10} \times \frac{2}{5} = \frac{1}{25}$
2, 2	2	$\frac{1}{10} \times \frac{1}{10} = \frac{1}{100}$

Distribuição de $\overline{X}$.

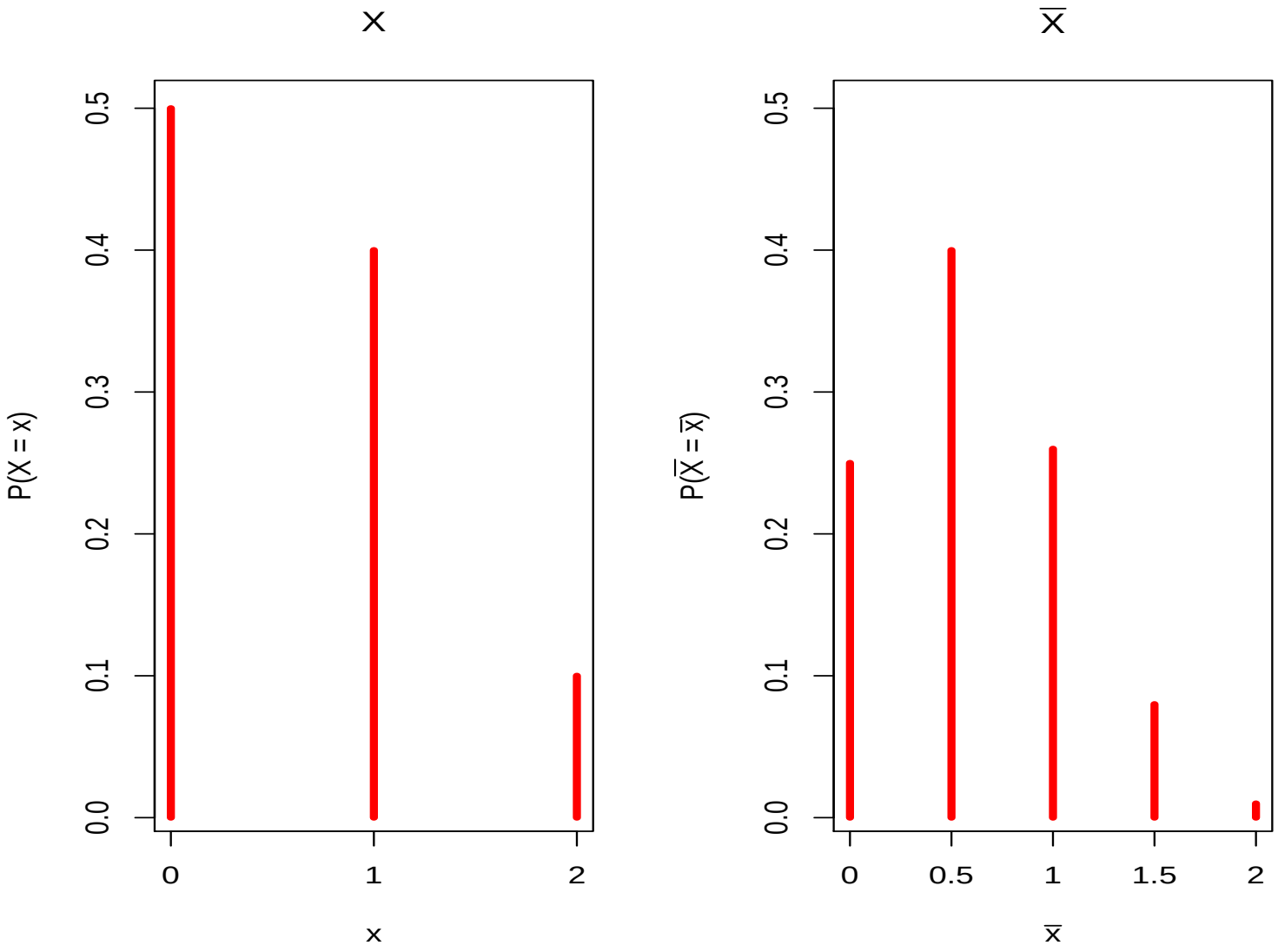
$\overline{x}$	$\mathsf{P}(\overline{X} = \overline{x})$
0	$\frac{1}{4}$
1/2	$\frac{1}{5} + \frac{1}{5} = \frac{2}{5}$
1	$\frac{1}{20} + \frac{4}{25} + \frac{1}{20} = \frac{13}{50}$
3/2	$\frac{1}{25} + \frac{1}{25} = \frac{2}{25}$
2	$\frac{1}{10} \times \frac{1}{10} = \frac{1}{100}$
Total	1

Exercício

(a) Calcule $E(X)$ e $E(\overline{X})$. Surpresa?

(b) Resolva o exercício com $n = 3$.

Distribuições de X e $\overline{X}$



Aproximação de Monte Carlo

Geração de um grande número de amostras de tamanho $n = 2$ da distribuição de X.
Para cada amostra gerada calculamos $\overline{x} = (x_1 + x_2)/2$.
A tabela de frequências relativas de $\overline{x}$ é uma aproximação da distribuição de X.

Aproximação de Monte Carlo

Solução em R

```
## Aproximação de Monte Carlo
# Distribuição de X
x <- 0:2
fx <- c(1 / 2, 2 / 5, 1 / 10)
# Número de repetições
M <- 10000

# Cálculo das médias
xb <- c()
for (m in 1:M) {
 xb[m] <- mean(sample(x, 2, replace = TRUE, prob = fx))
}
print(table(xb) / M)
0         0.5       1   1.5     2
0.2491 0.3977 0.2608 0.0814 0.0110
```

Média amostral

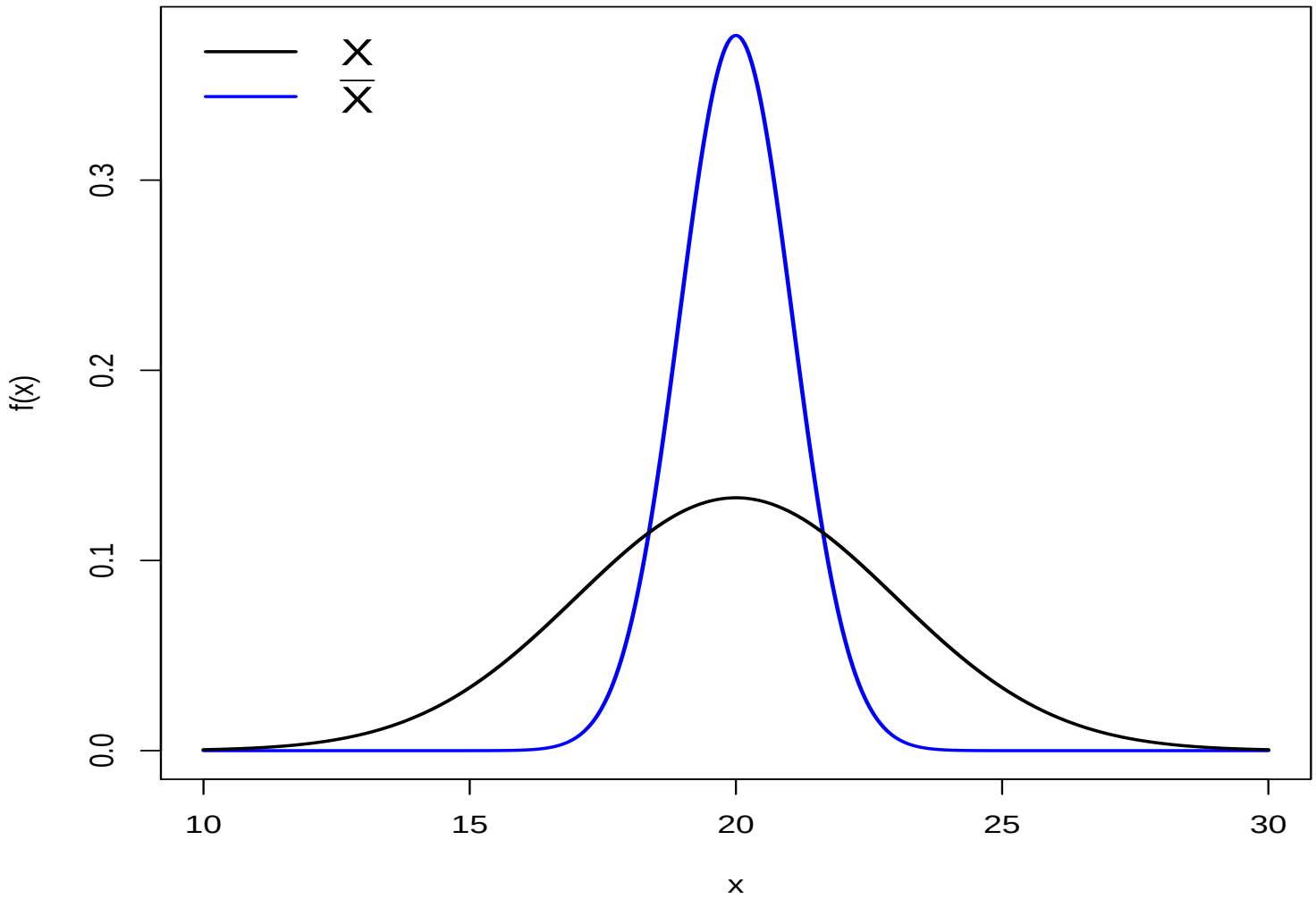
Propriedade. A v.a. $Z = \{X - E(X)\}/\sqrt{Var(X)}$ tem média 0 e variância 1 (padronização).

Propriedade. Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de uma população com média μ e variância σ^2, então $\overline{X}$ tem média μ e variância σ^2/n.

Portanto, $\sqrt{n}(\overline{X} - \mu)/\sigma$ tem média 0 e variância 1.

Propriedade. Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de uma população $\mathsf{N}(\mu, \sigma^2)$, então $\overline{X} \sim \mathsf{N}(\mu, \sigma^2/n)$.

Média amostral



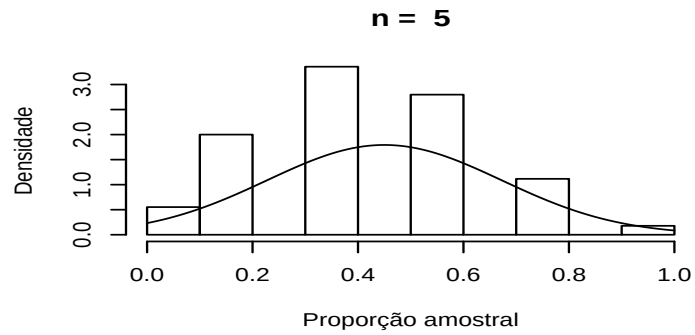
Média amostral

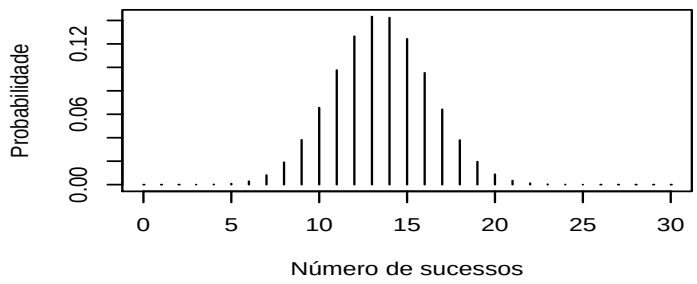
Teorema central do limite. Se $X_1, X_2, \ldots, X_n$ é uma amostra aleatória de uma população com média μ e variância σ^2 $(0 < \sigma^2 < \infty)$, então $\overline{X} \sim \mathsf{N}(\mu, \sigma^2/n)$, aproximadamente.
Quanto maior n, melhor a aproximação.
X pode ser discreta ou contínua.

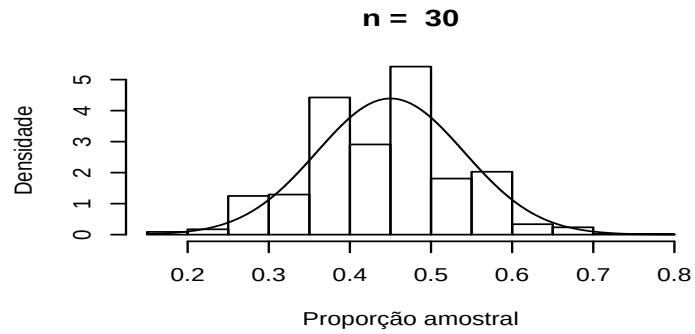
Exemplo

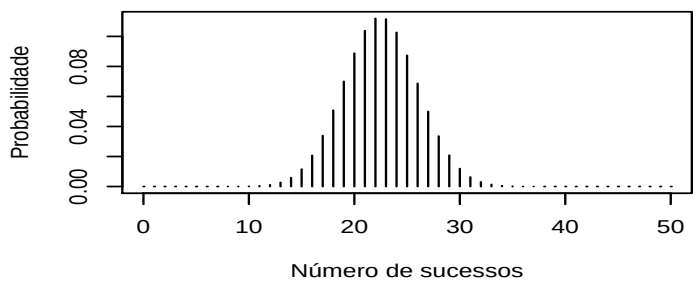
População binomial com parâmetros n e 0,45.
Experimento. Amostras com diferentes n são selecionadas (função `rbinom` em R) e os histogramas são obtidos (função `hist`).
Aproximação pela distribuição $N(0,45; \frac{1}{n \times 0{,}45 \times (1-0{,}45)})$ (funções `curve` e `dnorm`).

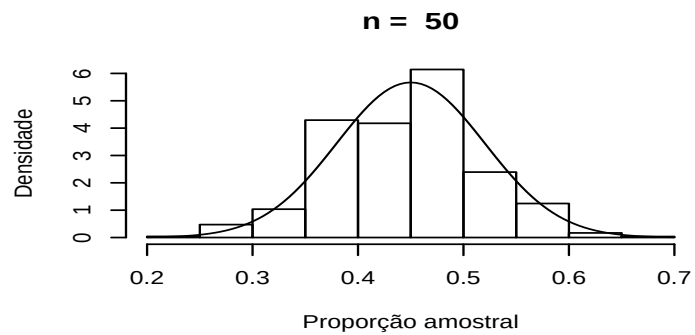
Média amostral









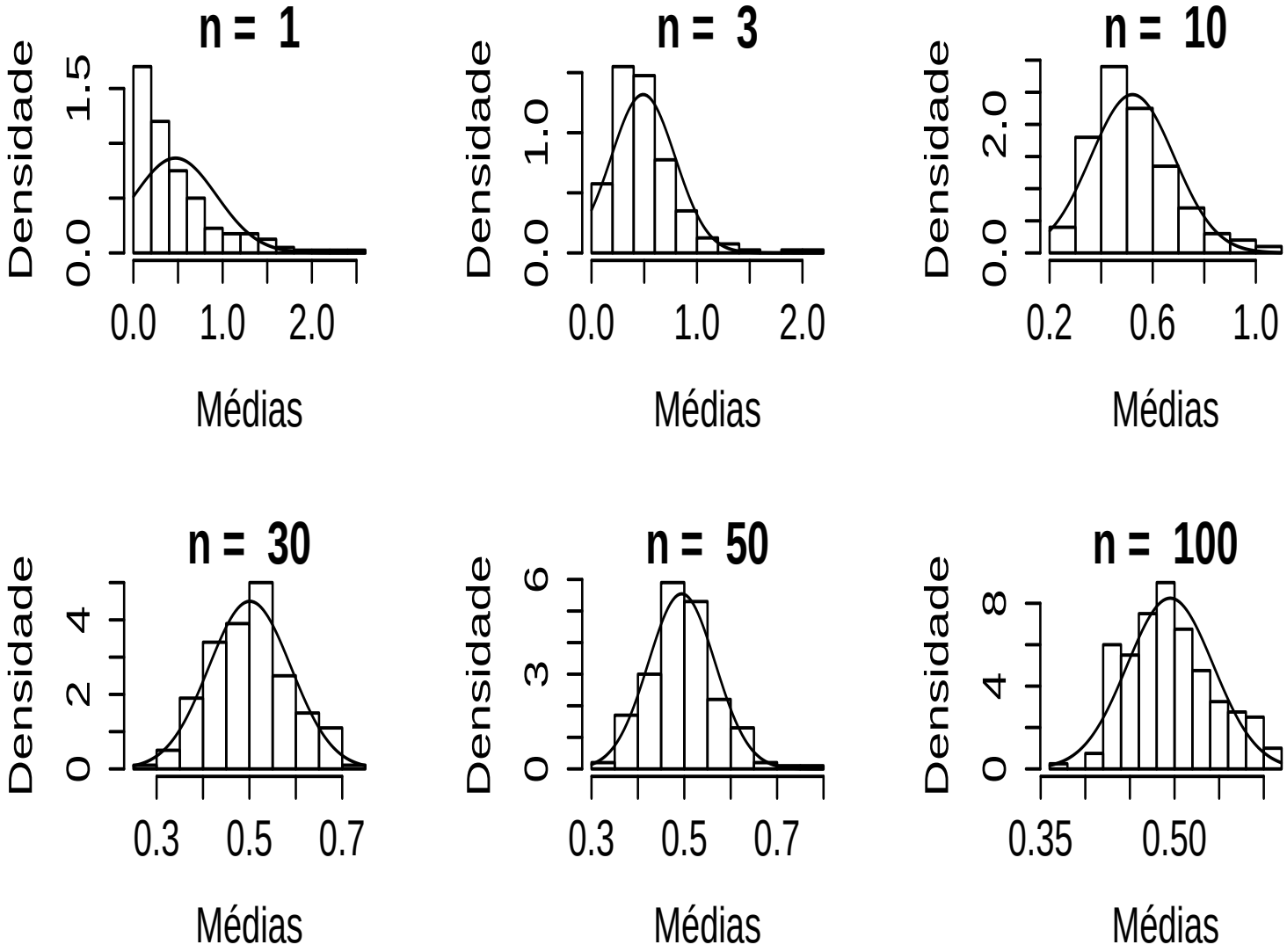


Média amostral

Exemplo

População exponencial com média 1/2 e variância 1/4.
Experimento. Amostras com diferentes n são selecionadas (função `rexp` em R) e os histogramas são obtidos (função `hist`).
Aproximação pela distribuição $N(\frac{1}{2}, \frac{1}{4n})$ (funções `curve` e `dnorm`).

Média amostral



Exemplo

De uma população com média 10 e variância 7 selecionou-se uma amostra aleatória, $X_1, X_2, \ldots, X_5$. *Calcular* $P(\overline{X} > 11,5)$.
A distribuição de X *não foi informada.*
Padronização: $Z = \sqrt{5}(\overline{X} - 10)/\sqrt{7}$ *tem média 0 e variância 1.*
Calculamos

$$P(\overline{X} > 11,5) = P\left(\frac{\sqrt{5}(\overline{X} - 10)}{\sqrt{7}} > \frac{\sqrt{5}(11,5 - 10)}{\sqrt{7}}\right) = P(Z > 1,268).$$

Sabemos que $Z \sim N(0,1)$, *aproximadamente.*
Solução em R. `1 - pnorm(sqrt(5) * 1.5 / sqrt(7))` $= 0.1024469$.
Uma tabela da distribuição $N(0,1)$ *poderia ser consultada. Se a distribuição de* X *for normal, a solução apresentada é exata.*

Proporção amostral

Um experimento admite apenas um de dois possíveis resultados: sucessso e insucesso (experimento de Bernoulli).
Definimos a v.a. X tal que $X = 1$ se ocorre sucesso e $X = 0$, caso contrário. Ocorre sucesso com probalidade p.
O experimento é repetido n vezes de forma independente. Definimos $Y = \sum_{i=1}^{n} X_i$, que representa o número de sucessos na amostra de tamanho n.
$\overline{X} = Y/n$ representa a proporção amostral de sucessos, com valores no conjunto $\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$ (v.a. discreta).

Proporção amostral

Como $X \sim$ Bernoulli(p), então $Y \sim$ binomial(n, p). Portanto, se $k \in \{0, 1, 2, \ldots, n-1, n\}$, temos que

$$\mathsf{P}\left(\overline{X} = \frac{k}{n}\right) = P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

ou seja, a distribuição da proporção amostral é exata.
Por outro lado,

$$\overline{X} \sim \mathsf{N}\left(p, \frac{p(1-p)}{n}\right), \text{ aproximadamente.}$$

Exemplo

Um esquema de controle de qualidade foi planejado de modo a garantir um máximo de 10% de itens defeituosos produzidos. Periodicamente seleciona-se uma amostra de 20 itens e, havendo mais de 15% de itens defeituosos, a produção é interrompida.
Calcule a probabilidade de uma interrupção desnecessária.

X_i indica a situação do item, em que $X_i = 1$ indica um item defeituoso (evento "sucesso"), $i = 1, \ldots, n$, com $n = 20$.
Supondo uma amostra aleatória, $Y = \sum_{i=1}^{20} X_i \sim$ binomial$(20, p)$ (número de itens defeituosos em uma amostra de $n = 20$ itens).

Exemplo

No pior caso[1] temos $p = 10\% = 0,1$. Assim, $Y \sim \text{binomial}(20; 0,1)$. Calculamos

$$\mathsf{P}(\overline{X} > 0,15) = \mathsf{P}(\overline{X} > 3/20) = \mathsf{P}(Y > 3)$$

$$= \mathsf{P}(Y \geqq 4) = 1 - \mathsf{P}(Y < 4) = 1 - \mathsf{P}(Y \leqq 3)$$

$$= 1 - \sum_{j=0}^{3} \binom{20}{j} 0,1^j \times 0,9^{20-j} = 0,1330.$$

Em R: `1 - pbinom(3, 20, 0.1)` ou `pbinom(3, 20, 0.1, lower.tail = FALSE)`.

[1] Pode ser provado que a probabilidade de uma interrupção desnecessária cresce com p.

Exemplo

Solução aproximada. Sabemos que

$$\overline{X} \sim \mathsf{N}\left(0,1; \frac{0,1 \times 0,9}{20}\right), \text{ aproximadamente.}$$

Calculamos

$$\begin{aligned}\mathsf{P}(\overline{X} > 0,15) &= \mathsf{P}\left(\frac{\sqrt{20}(\overline{X} - 0,1)}{\sqrt{0,1 \times 0,9}} > \frac{\sqrt{20}(0,15 - 0,1)}{\sqrt{0,1 \times 0,9}}\right) \\ &\cong \mathsf{P}(Z > 0,745356) = 1 - \mathsf{P}(Z \leqq 0,745356),\end{aligned}$$

em que $Z \sim \mathsf{N}(0,1)$. Consultando uma tabela de Z obtemos 0,2280.
A aproximação é satisfatória? Realize um experimento para ajudar em sua resposta.
Em R: `z0 <- sqrt(20) * (0.15 - 0.1) / sqrt(0.1 * 0.9)`
`1 - pnorm(z0)` ou `pnorm(z0, lower.tail = FALSE)`.