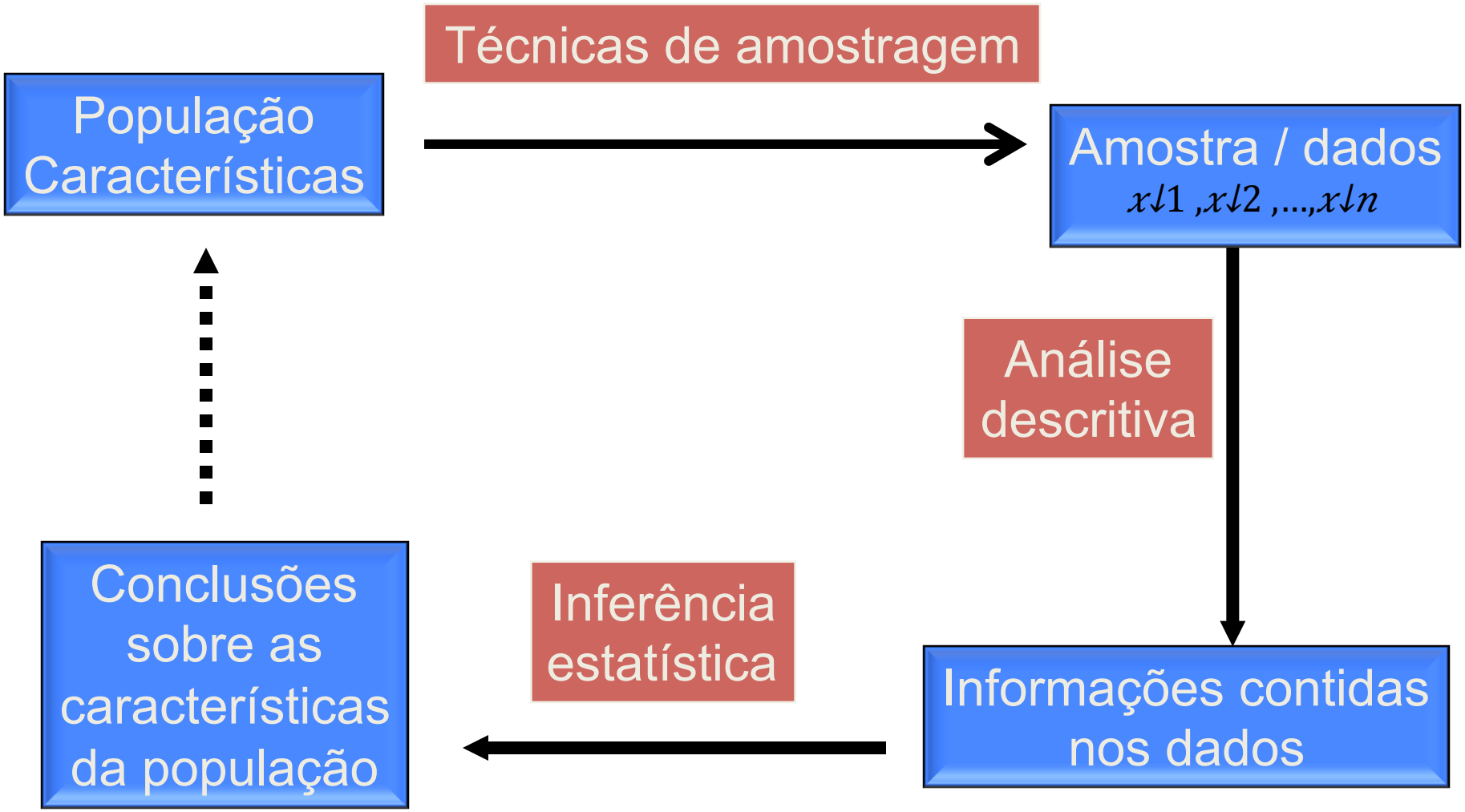
Aula 1. Revisões de MAE0219

Estatística

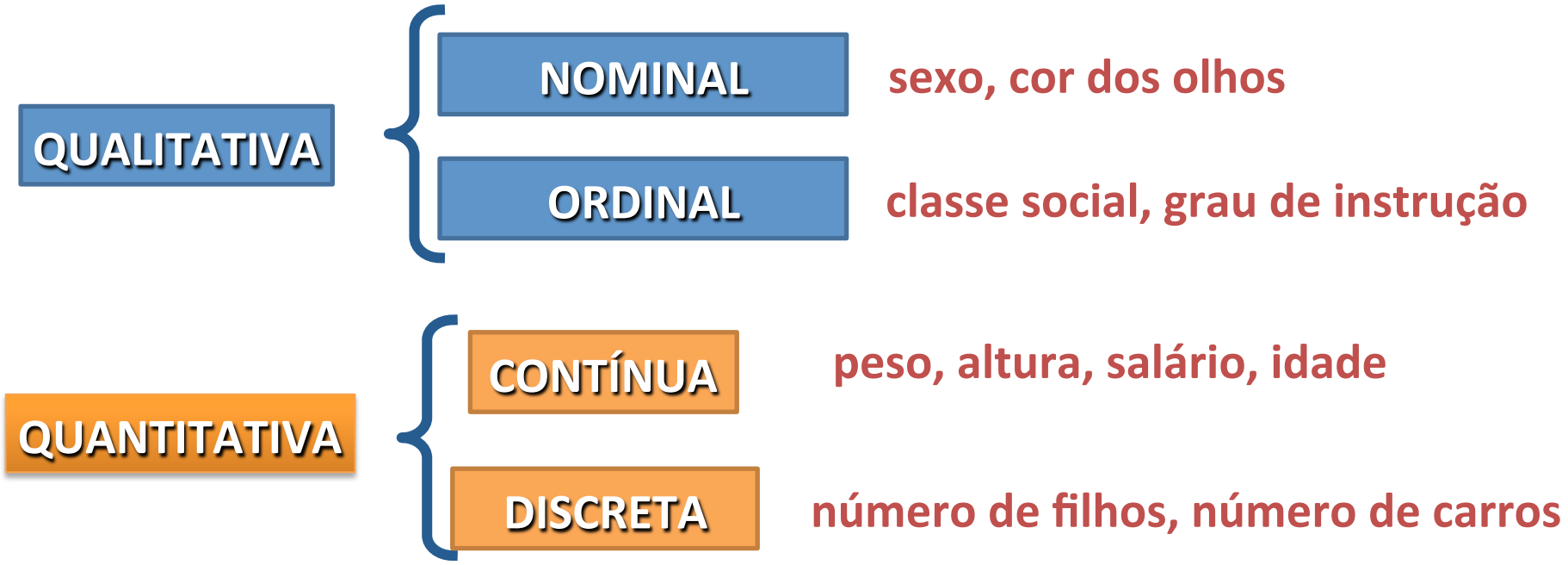


Técnicas de amostragem

Análise descritiva = resumo de dados

qualquer característica associada a uma população chamamos de variável aleatória

classificação de variáveis aleatórias



Resumo de variáveis quantitativas

MEDIDAS DE POSIÇÃO

Mínimo, Máximo, Moda, Média, Mediana, Percentis.

MEDIDAS DE DISPERSÃO

Amplitude, Intervalo-Interquartil, Variância, Desvio Padrão, Coeficiente de Variação.

Medidas de Posição

- **Máximo (max):** a maior observação.
- **Mínimo (min):** a menor observação.
- **Moda (mo):** é o valor (ou atributo) que ocorre com maior frequência.

Dados: 4, 5, 4, 6, 5, 8, 4

max = 8 min = 4 mo = 4

•Média

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Dados: 2, 5, 3, 7, 8

$$\bar{x} = \frac{2+5+3+7+8}{5} = 5$$

- **Mediana**

A mediana é o valor da variável que ocupa a posição central de um conjunto de *n* dados ordenados.

Posição da mediana: $\frac{n+1}{2}$

Exemplos

Dados: 2, 6, 3, 7, 8 $\Rightarrow n = 5$ (ímpar)

Dados ordenados: 2 3 6 7 8 $\Rightarrow \frac{5+1}{2} = 3 \Rightarrow Md = 6$

Posição da Mediana ↑

Dados: 4, 8, 2, 1, 9, 6 $\Rightarrow n = 6$ (par)

Dados ordenados: 1 2 4 6 8 9 $\Rightarrow \frac{6+1}{2} = 3{,}5$

↑

Md

$Md = (4 + 6) / 2 = 5$

•Percentis

O percentil de ordem $p \times 100$ $(0 < p < 1)$, em um conjunto de dados de tamanho n, é o valor da variável que ocupa a posição $p \times (n + 1)$ do conjunto de dados ordenados.

Casos particulares

percentil 50 = mediana ou segundo quartil (Md);

percentil 25 = primeiro quartil (Q_1);

percentil 75 = terceiro quartil (Q_3);

percentil 10 = primeiro decil.

Dados: 1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7 $\Rightarrow n=10$

Posição de Md: $0{,}5(n+1)= 0{,}5\times 11= 5{,}5$ $\Rightarrow Md = (3 + 3{,}1)/2 = 3{,}05$

Posição de $Q1$: $0{,}25\ (11) = 2{,}75$ $\Rightarrow Q_1=(\ 2+2{,}1)/2=2{,}05$

Posição de $Q3$: $0{,}75\ (11) = 8{,}25$ $\Rightarrow Q_3=(3{,}7+6{,}1)/2=4{,}9$

$Md = 3{,}05$ $Q_1 = 2{,}05$ $Q_3 = 4{,}9$

Dados: 0,9 1,0 1,7 2,9 3,1 5,3 5,5 12,2 12,9 14,0 33,6

$\Rightarrow n=11$

$Md = 5{,}3$ $Q1 = 1{,}7$ $Q3 = 12{,}9$

Medidas de Dispersão

Finalidade: encontrar um valor que resuma a variabilidade de um conjunto de dados.

- **Amplitude**

$$A = max - min$$

Para os grupos anteriores, temos:

Grupo 1, $A = 4$

Grupo 2, $A = 8$

Grupo 3, $A = 0$

•Intervalo-Interquartil

É a diferença entre o terceiro quartil e o primeiro quartil, ou seja, $Q3 - Q1$.

Dados: 1,9 2,0 2,1 2,5 3,0 3,1 3,3 3,7 6,1 7,7

$Q1 = 2,05$ e $Q3 = 4,9$

$Q3 - Q1 = 4,9 - 2,05 = 2,85$

- Variância

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

$$= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \sum_{1=1}^{n} \frac{x_i^{\,2}}{n-1} - \frac{n}{n-1} \bar{x}^{-2}$$

- Desvio padrão

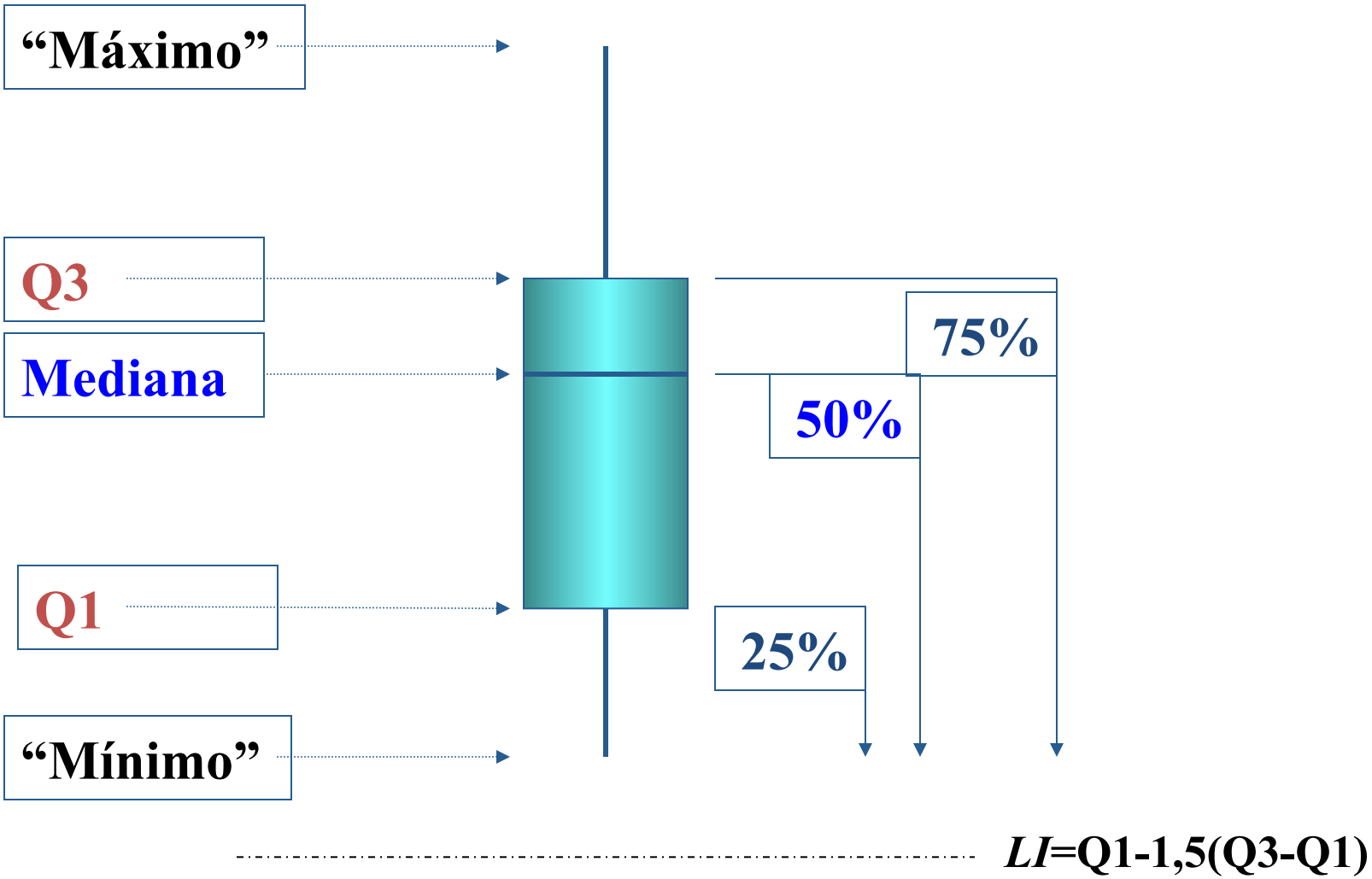
$$s = \sqrt{s^2}$$

•Coeficiente de Variação

- é uma medida de dispersão relativa;
- elimina o efeito da magnitude dos dados;
- exprime a variabilidade em relação à média.

$$CV = \frac{s}{\overline{x}} \times 100\%$$

Boxplot



"Máximo" é o maior valor menor que *LS*;

"Mínimo" é o menor valor maior que *LI*.

Histograma

Agrupar os dados em intervalos de classes

(distribuição de frequências)

Bases iguais

Construir um retângulo para cada classe, com base igual ao tamanho da classe e *altura proporcional à frequência da classe (f).*

Bases diferentes

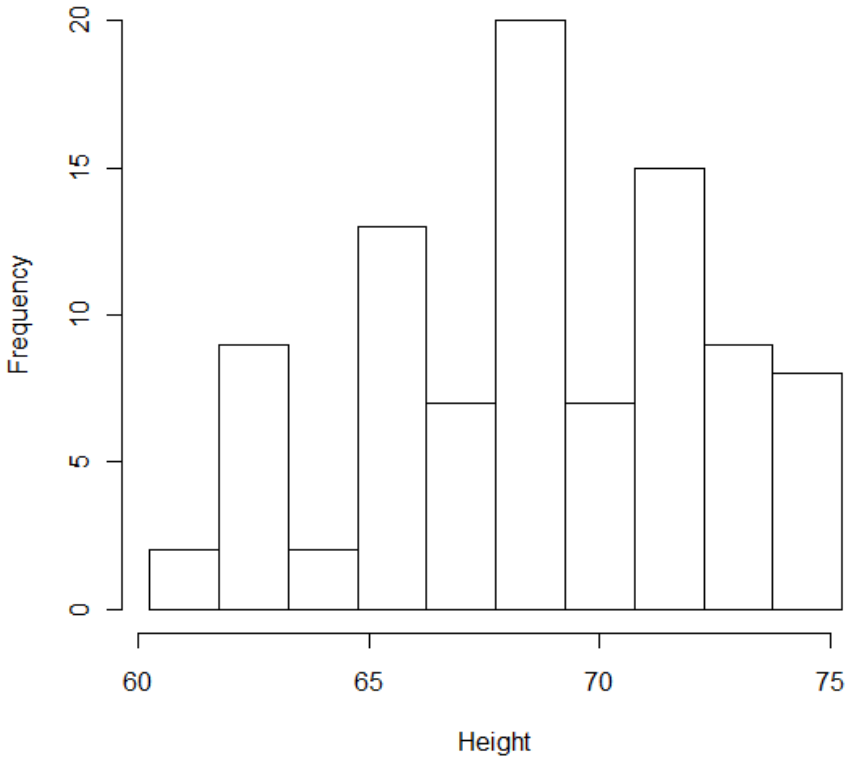
Construir um retângulo para cada classe, com base igual ao tamanho da classe e *área do retângulo igual a frequência relativa da classe(fr).* A altura será dada por

h = fr/base (densidade de frequência).

Arquivo *PULSE* – Histograma da altura (*Height*)

```
> b<-seq(60.25,75.25,by=1.50)
>hist(dados$Height,breaks=b,main=NULL,xlab="Height")
```

Classe de altura	f	fr
60,25 ⊢ 61,75	1	0,011
61,75 ⊢ 63,25	10	0,109
63,25 ⊢ 64,75	2	0,022
64,75 ⊢ 66,25	13	0,141
66,25 ⊢ 67,75	7	0,076
67,75 ⊢ 69,25	20	0,217
69,25 ⊢ 70,75	7	0,076
70,75 ⊢ 72,25	15	0,163
72,25 ⊢ 73,75	9	0,098
73,75 ⊢ 75,25	8	0,087
Total	92	1

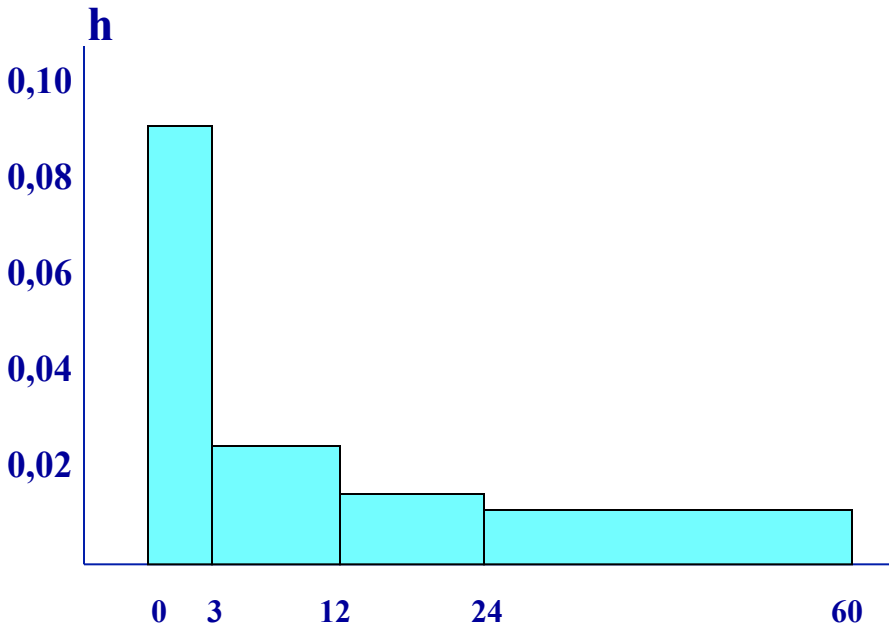


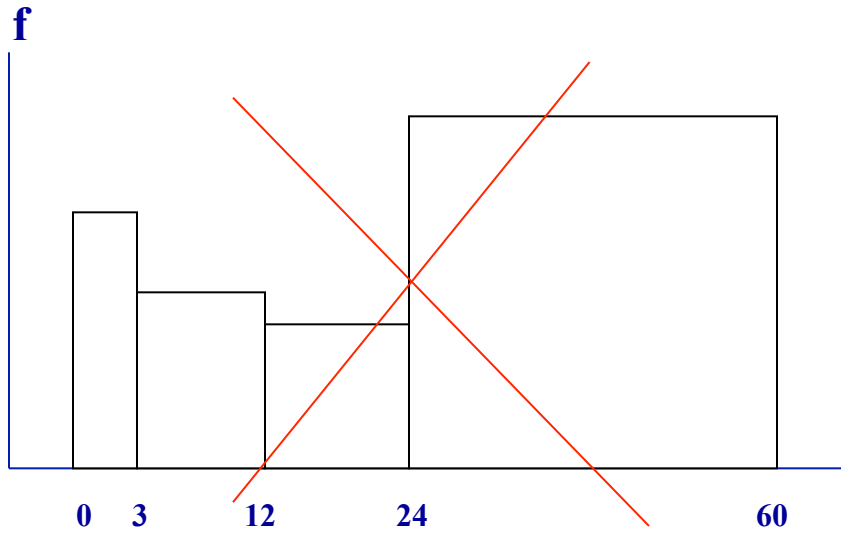
Exemplo: Classes desiguais

Distribuição das idades (em meses) de uma amostra de 500 crianças vacinadas

Classes (meses)	f	fr	h
0 |- 3	140	0,28	0,093
3 |- 12	100	0,20	0,022
12 |-24	80	0,16	0,013
24 |-60	180	0,36	0,010
Total	500	1,00	

h=fr/base





Distribuição de variável aleatória discreta.

Variável aleatória discreta e a sua distribuição podem ser definidas pela sua tabela

	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

onde todos os números $x↓i$ são diferentes e as probabilidades $p↓i$ de correspondentes valores satisfazem seguintes propriedades:

- $p↓i \geq 0$
- $p↓1 + p↓2 + ... + p↓n = 1$

Distribuição de variável aleatória discreta.

Variável aleatória X é número que sai em um experimento de jogada de um dado

	1	**2**	**3**	**4**	**5**	**6**
P	1/6	1/6	1/6	1/6	1/6	1/6

Variável aleatória X é soma dos números que saem em um experimento de jogada de dois dados

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
P	$1/36$	$2/36$	$3/36$	$4/36$	$5/36$	$6/36$	$5/36$	$4/36$	$3/36$	$2/36$	$1/36$

Distribuição de variável aleatória discreta.

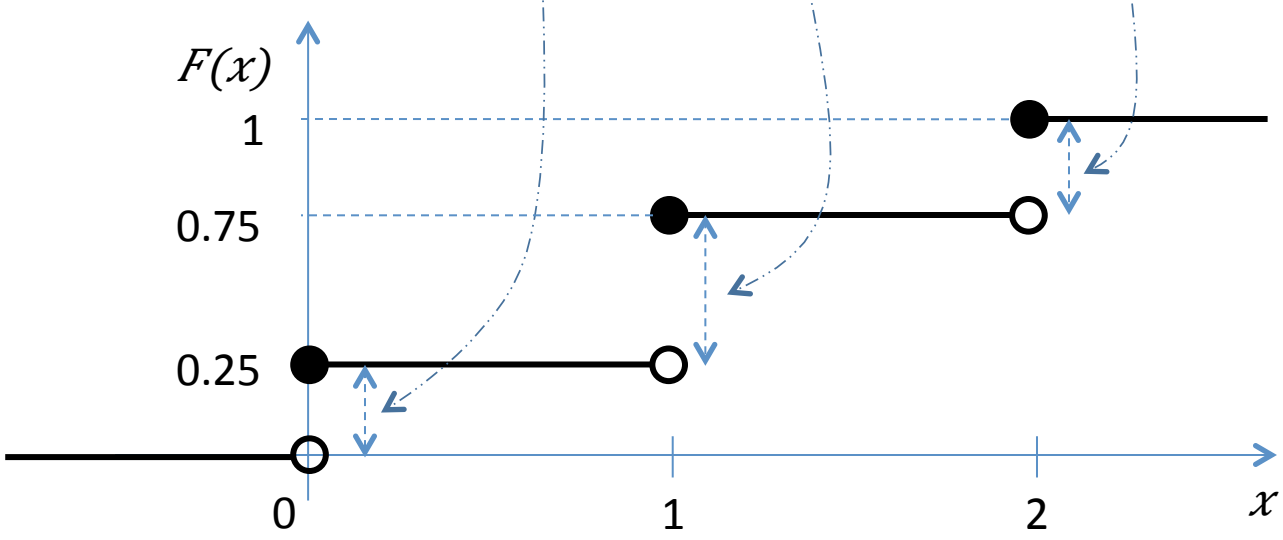
Outro jeito de apresentar uma variável aleatória discreta é função de distribuição cumulativa $F(x)$, ou, as vezes denotamos como $F\downarrow X\,(x)$ para destacar que uma função de variavel aleatoria X. Pela definição

$F(x)=P(X\leq x)$

Por exemplo, consideramos v.a. X dada pela tabela

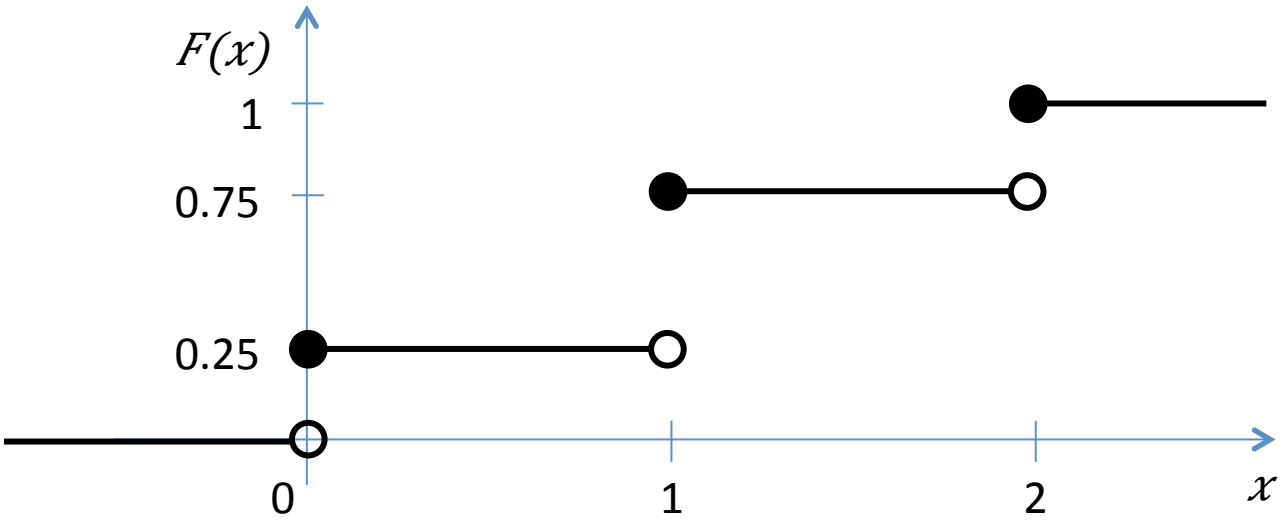
	0	**1**	**2**
P	0.25	0.5	0.25

Desenhamos gráfico de $F(x)$:



Distribuição de variável aleatória discreta.

	0	1	2
P	0.25	0.5	0.25



Distribuição de variável aleatória discreta.

Distribuição Bernoulli.

Supomos um simples modelo de alteração de preço de uma ação. Seja $s\downarrow 1$ o preço no instante "agora". No próximo instante (um tick, próxima negociação, próximo dia etc.) o preço aumentou com probabilidade p ou diminuiu em um ponto com probabilidade $q=1-p$. Se o evento "preço aumentou" vou codificar como "1" e o evento "preço diminuiu" como "0", então tenho uma variável Bernoulli

	0	**1**
P	q	p

Caso quero a distribuição de incremento do preço posso considerar

	-1	**1**
P	q	p

Distribuição de variável aleatória discreta.

	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

$$E(X) = \sum_{i=1}^{n} x_i p_i$$

Variância: É o valor esperado da v.a. $(X - E(X))^2$, ou seja, se X assume os valores $x_1, x_2, \ldots, x_n$, então

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

Notação: $\sigma^2 = \mathrm{Var}(X)$.

Da relação acima, segue que

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

Desvio Padrão: É definido como a raiz quadrada positiva da variância, isto é,

$$\mathrm{DP}(X) = \sqrt{\mathrm{Var}(X)}.$$

Notação: $\sigma = \mathrm{DP}(X)$.

Distribuição de variável aleatória discreta. Propriedades Esperança e Variância.

$E(X+a)=E(X)+a$
$E(aX)=aE(X)$
$E(a)=a$

$Var(X+a)=Var(X)$
$Var(aX)=a\uparrow 2\ Var(X)$
$Var(a)=0$

Para duas v.a. quaisquer X,Y
$E(X+Y)=E(X)+E(Y)$

Para duas v.a. quaisquer X,Y
e independentes
$Var(X+Y)=Var(X)+Var(Y)$

Observação: Seja $Y=f(X)$
em geral $E(Y)\neq f(E(X))$, mas isso é verdade, caso f é uma função linear

Distribuição binomial:

A v.a. X correspondente ao **número de sucessos em n ensaios de Bernoulli independentes e com mesma probabilidade p de sucesso** tem *distribuição binomial com parâmetros n e p.*

Sua função de probabilidade é dada por

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad , \quad k = 0, 1, \ldots, n.$$

Notação: $X \sim B(n; p)$.

Resultado: Se $X \sim B(n; p)$, então

média: $\mu = \mathrm{E}(X) = np$

variância: $\sigma^2 = \mathrm{Var}(X) = np(1-p) = npq$

Variáveis Aleatórias Contínuas

Função Densidade de Probabilidade

A função densidade de probabilidade (f.d.p.) de uma variável aleatória X é uma função $f(x) \geq 0$ cuja área total sob a curva seja igual à unidade. Em termos matemáticos

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Variáveis Aleatórias Contínuas

Probabilidade de Eventos

A probabilidade $P(a \leq X \leq b)$ corresponde à área sob a curva no intervalo $[a, b]$. Em termos matemáticos

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

Esperança

Definição

A esperança matemática de uma variável aleatóra contínua X fica dada por

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{+\infty} xf(x)dx.$$

Variância

Definição

A variância de uma variável aleatória X contínua é definida por

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}[X - \mathrm{E}(X)]^2 \\ &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2, \end{aligned}$$

em que

$$\mathrm{E}(X^2) = \int_{-\infty}^{+\infty} x^2 f(x) dx.$$

Função de Distribuição Acumulada

Definição

A Função de Distribuição Acumulada de uma variável aleatória T contínua é definida por

$$F(x) = P(T \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Distribuição Normal

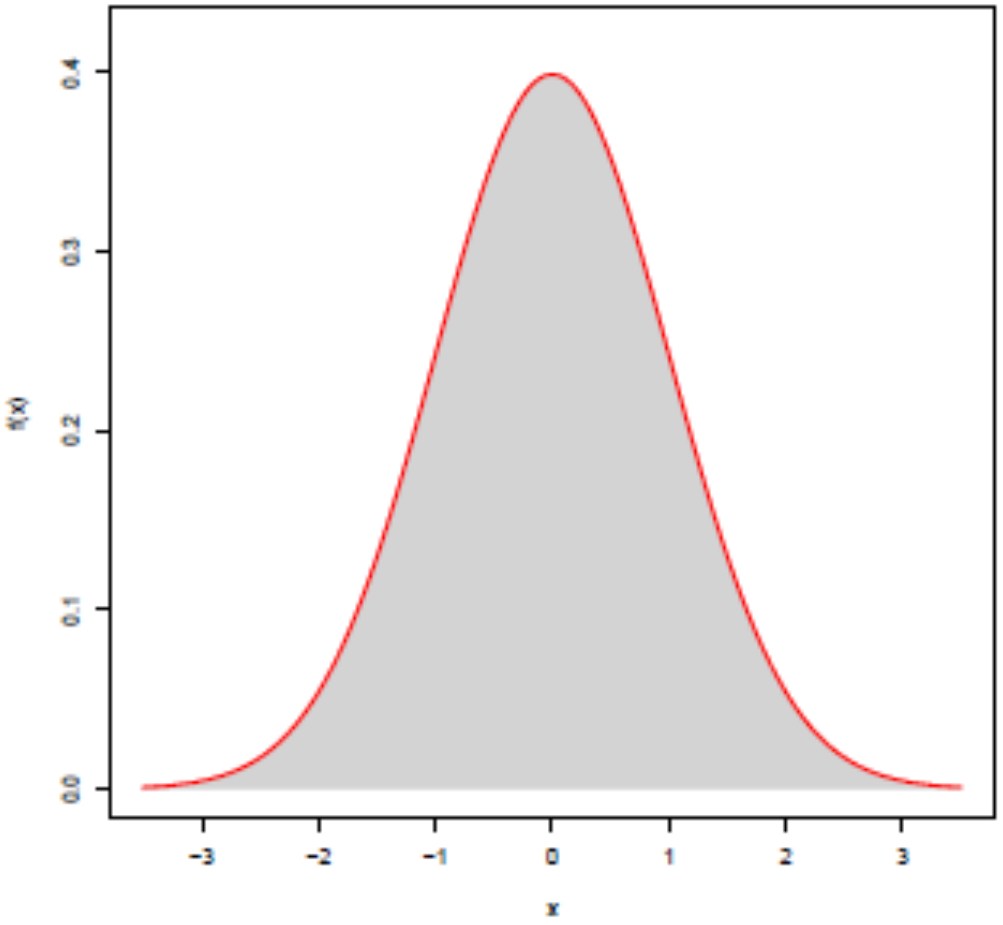
Distribuição Normal

Se X é uma variável aleatória com distribuição normal de média μ e variância σ^2, a função densidade de probabilidade de X é definida por

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

para $-\infty < x, \mu < +\infty$ e $\sigma > 0$. Notação: $X \sim \mathrm{N}(\mu, \sigma^2)$.

Descrição de $f(x)$ de uma N(0,1).



Distribuição Normal

Padronização

Se $X \sim \mathrm{N}(\mu, \sigma^2)$ e $Z \sim \mathrm{N}(0, 1)$ (normal padrão), então

$$P(X \leq x) = P\left(Z \leq \frac{x - \mu}{\sigma}\right),$$

ou seja, todos os cálculos podem ser feitos pela normal padrão.

Distribuição Normal : Valores de $P(Z \leq z) = A(z)$

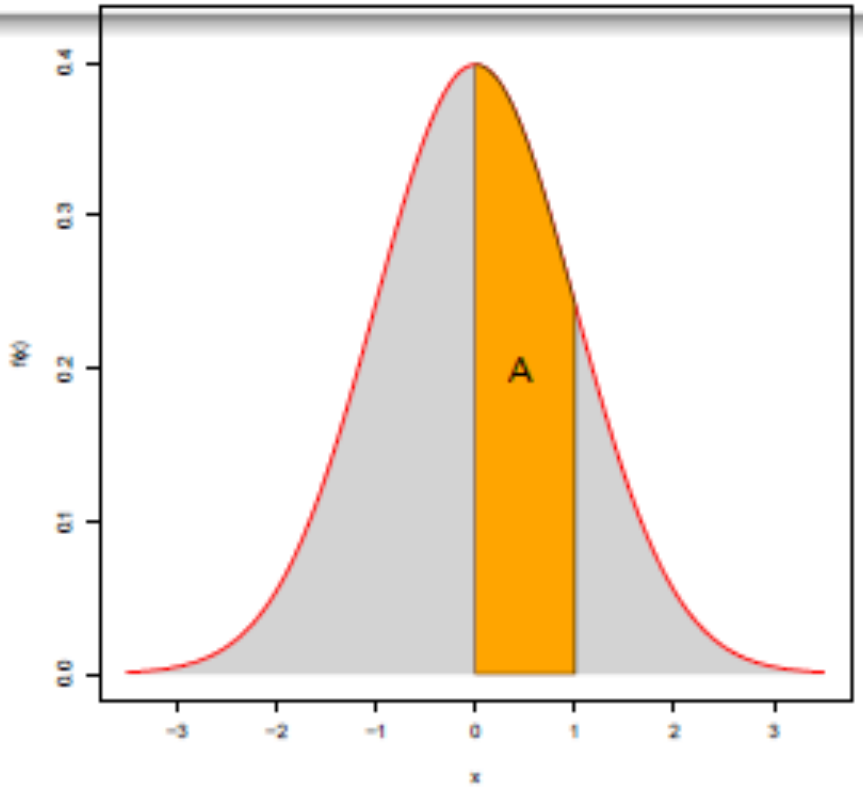
Parte inteira e primeira decimal de z	Segunda decimal de z									
	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Distribuição N(0,1)

Cálculo de Probabilidades

Por exemplo, a probabilidade $A = P(0 \leq X \leq 1)$ pode ser calculada pela diferença

$$P(X \leq 1) - P(X \leq 0) = 0,841 - 0,5 = 0,341.$$

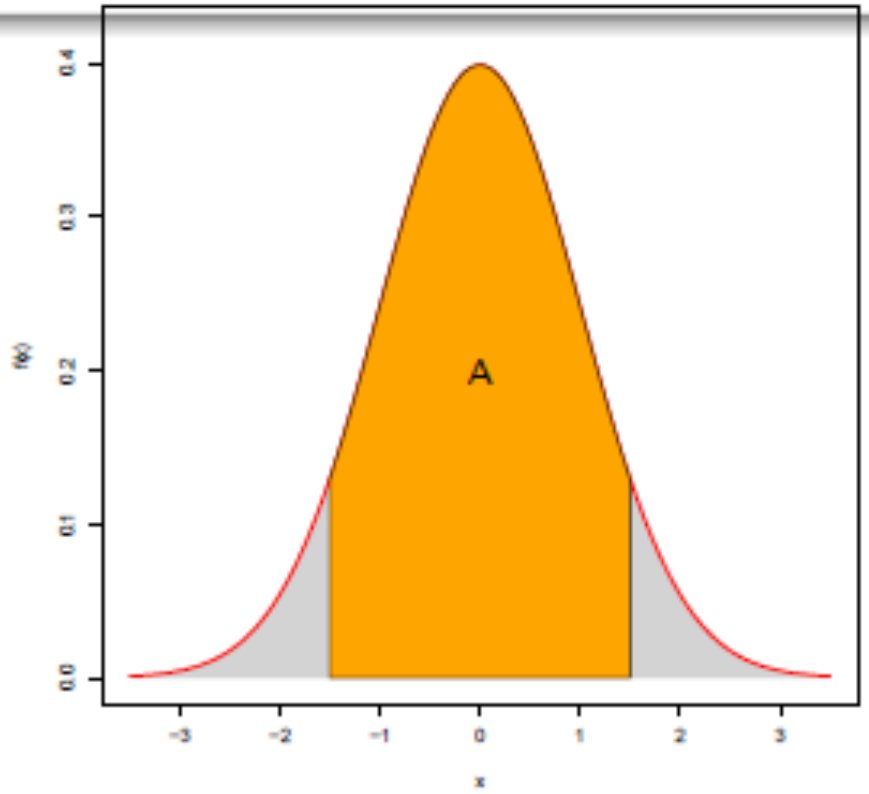


Distribuição N(0,1)

Cálculo de Probabilidades

Para calcular a probabilidade $A = P(-1 \leq X \leq 1)$ podemos usar o fato da distribuição ser simétrica na média. Assim,

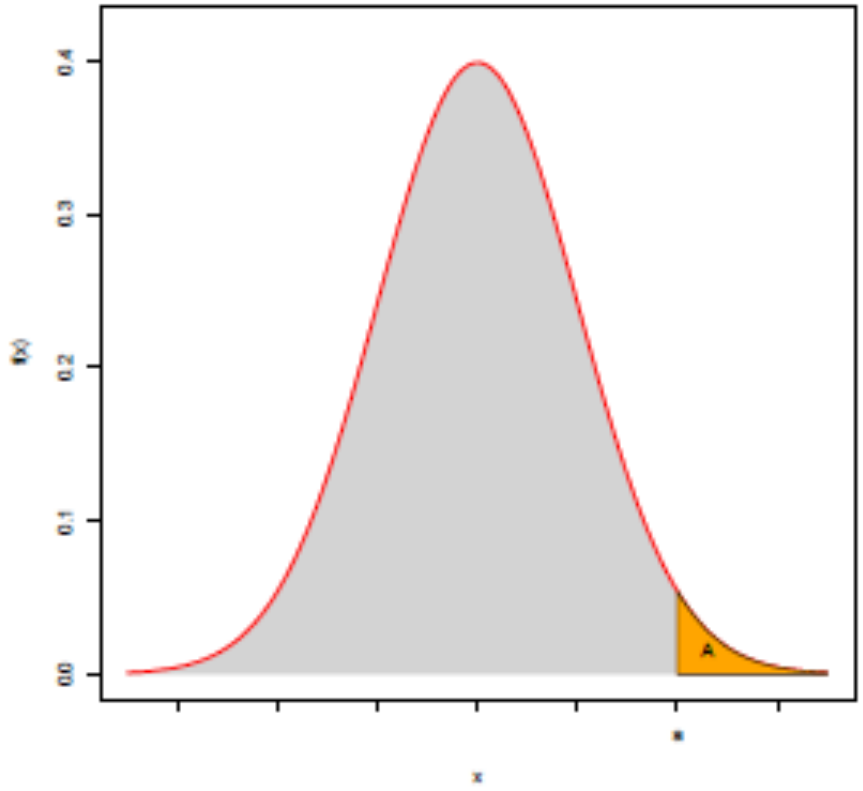
$$P(-1 \leq X \leq 1) = 2 \times P(0 \leq X \leq 1) = 2 \times 0,341 = 0,682.$$



Distribuição N(0,1)

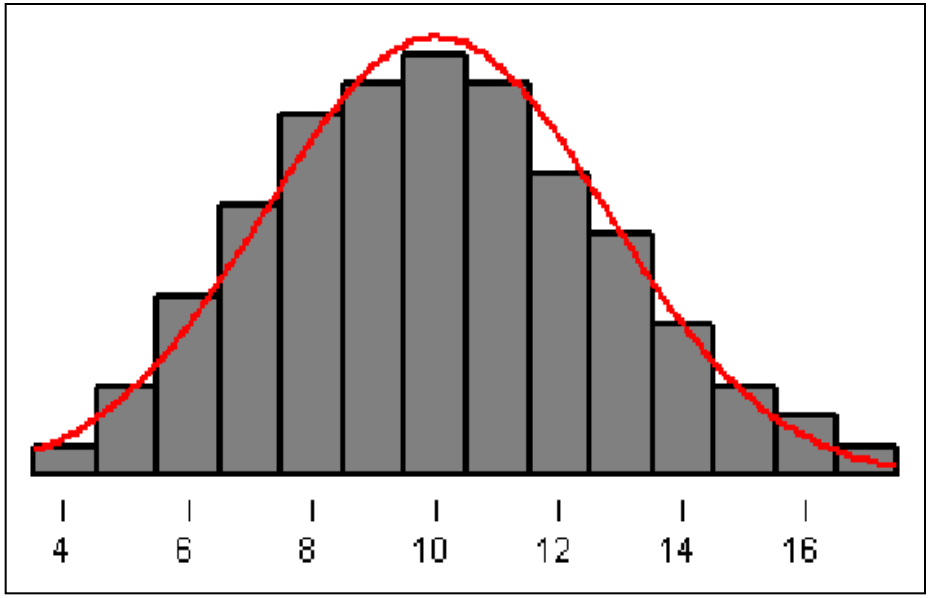
Cálculo de Percentil

Como encontrar a tal que $A = P(X \geq a)$, em que $A = 0,98$? Pelo R podemos encontrar $a = 2,054$ usando o comando qnorm(0.98).



Aproximação da binomial pela normal

Considere a binomial com $n = 50$ e $p = 0,2$, representada pelo histograma



$P(Y = 13)$ é igual a área do retângulo de base unitária e altura igual a $P(Y = 13)$; similarmente, $P(Y = 14)$, etc...

Logo, $P(Y \geq 13)$ é igual à soma das áreas dos retângulos correspondentes.

A idéia é aproximar tal área pela área sob uma curva normal, à direita de 13.

→ *Qual curva normal?*

$X \sim b(n\,;\,p) \quad \Rightarrow \quad E(X) = np$
$Var(X) = np(1-p)$

Parece razoável considerar a normal com média e variância iguais às da binomial, ou seja, aproximamos a distribuição de probabilidades de *X* pela distribuição de probabilidades de uma variável aleatória *Y*, sendo

$Y \sim N(\mu_y\,;\,\sigma_y^2)$ com $\mu_y = np$ e $\sigma_y^2 = np(1-p)$.

Portanto,

- $P(a \le X \le b) \approx P(a \le Y \le b)$
- $P(X \ge a) \approx P(Y \ge a)$
- $P(X \le b) \approx P(Y \le b)$

com $Y \sim N(np;\, np(1-p))$.

O cálculo da probabilidade **aproximada** é feito da forma usual para a distribuição normal:

$P(a \leq X \leq b) \approx P(a \leq Y \leq b)$ com $Y \sim N(np;\ np(1-p))$.

Lembrando que

$$Z = \frac{Y - np}{\sqrt{np(1-p)}} \sim N(0;1),$$

então

$$P(a \leq Y \leq b) = P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq \frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{b - np}{\sqrt{np(1-p)}} \right)$$

$$= P\left(\frac{a - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - np}{\sqrt{np(1-p)}} \right).$$

Observações :

1 - A aproximação da distribuição binomial pela normal é boa quando $np(1\text{-}p) \geq 3$.

2 - A demonstração da validade desta aproximação é feita utilizando-se o Teorema Central do Limite (TCL).

Teorema do Limite Central

Enunciado para a Soma Amostral

Para variáveis aleatórias $X_1, \ldots, X_n$ independentes e com mesma distribuição de média μ e variância σ^2 finitas, a distribuição da soma

$$X = X_1 + \cdots + X_n$$

se aproxima à medida que n cresce da distribuição de $Y \sim \mathrm{N}(\mu_X, \sigma_X^2)$, em que $\mu_X = n\mu$ e $\sigma_X^2 = n\sigma^2$.

Teorema do Limite Central

Aproximação para n Grande

$$\begin{aligned} P(a \leq X \leq b) &\cong P(a \leq Y \leq b) \\ &= P\left(\frac{a - n\mu}{\sigma\sqrt{n}} \leq Z \leq \frac{b - n\mu}{\sigma\sqrt{n}}\right), \end{aligned}$$

em que $Z \sim \mathrm{N}(0, 1)$.

Teorema do Limite Central

Média Amostral

Para a média amostral $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ temos que

$$\begin{aligned} \mathrm{E}(\bar{X}) &= \frac{\mathrm{E}(X_1) + \cdots + \mathrm{E}(X_n)}{n} \\ &= \frac{n\mu}{n} = \mu \quad \text{e} \\ \mathrm{Var}(\bar{X}) &= \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2} \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Teorema do Limite Central

Enunciado para a Média Amostral

Para variáveis aleatórias $X_1, \ldots, X_n$ independentes e com mesma distribuição de média μ e variância σ^2 finitas, a distribuição da média amostral

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

se aproxima à medida que n cresce da distribuição de $Y \sim \mathrm{N}(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$, em que $\mu_{\bar{X}} = \mu$ e $\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$.