Técnicas Computacionais em Probabilidade e Estatística I

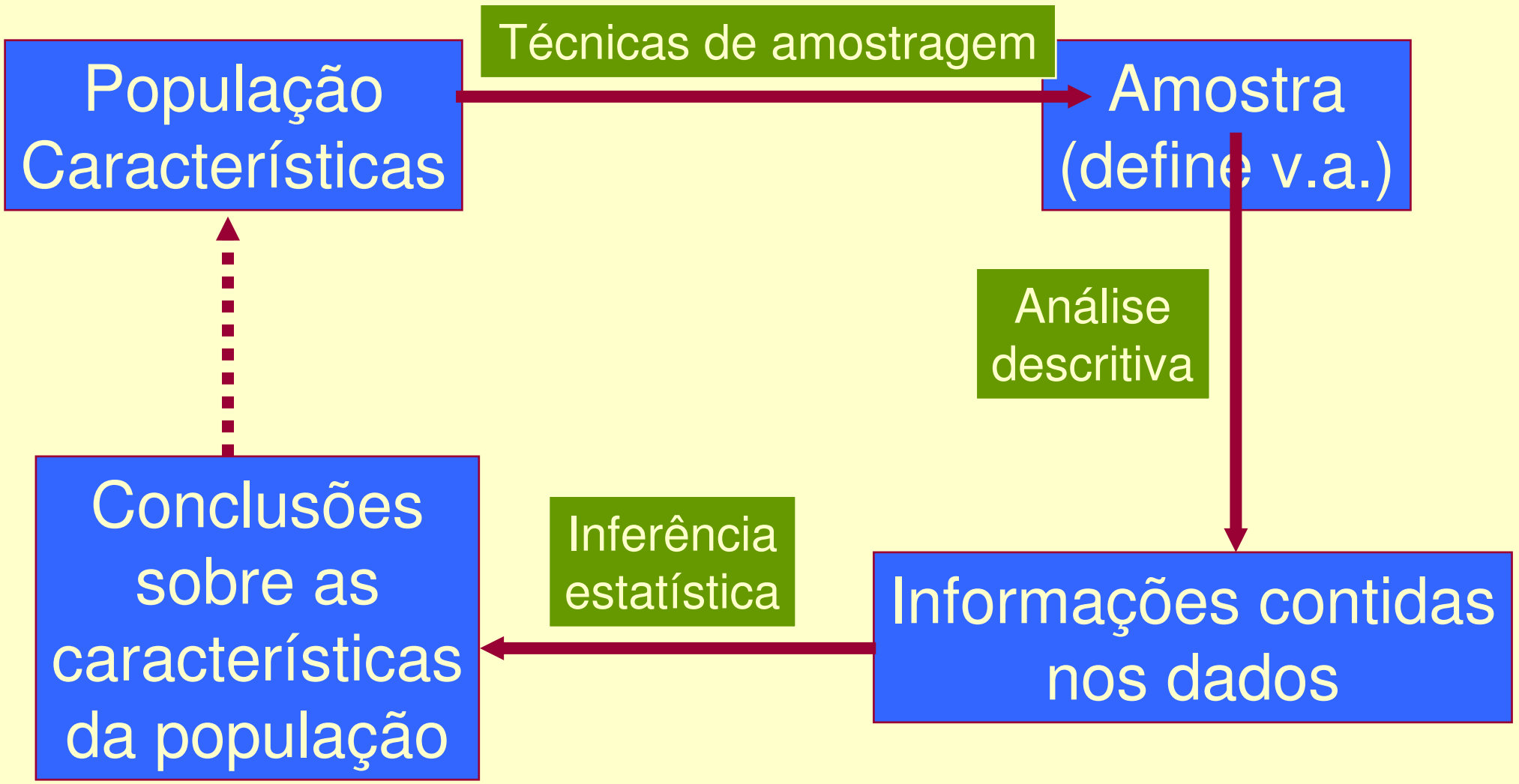
Aula I

Chang Chiann
MAE 5704- IME/USP
1º Sem/2008

Análise de Um conjunto de dados

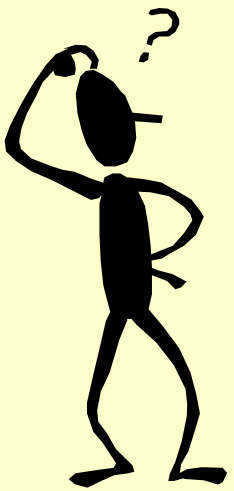
- objetivo: tratamento de um conjunto de dados.
- uma amostra de uma população.
- uma variável aleatória (v.a.) X, cuja distribuição de probabilidade define a população.

Estatística



Estatítica Descritiva

O que fazer com as observações que coletamos?



Primeira Etapa:

Resumo dos dados = Estatística descritiva

Variável:

Qualquer característica associada a uma população.

Classificação das variáveis:

- **QUALITATIVA**
  - **NOMINAL** — sexo, cor dos olhos
  - **ORDINAL** — classe social, grau de instrução
- **QUANTITATIVA**
  - **CONTÍNUA** — peso, altura, salário, idade
  - **DISCRETA** — número de filhos, número de carros

. Considere uma população com N indivíduos:

$$P = \{ 1, 2, ..., N\}$$

. v.a. X definida sobre cada um desses indivíduos, com valores distintos $\{x_1, ..., x_k\}$.

. $P_i = P(x_i) = P(X=x_i)$: a probabilidade de X assumir o valor x_i.

. Os pares (x_i, p_i), i=1, .., k, constituem a distribuição de probabilidades de X.

A média e a variância de X:

$$\mu = E(X) = \Sigma x_i p_i$$

$$\sigma^2 = Var(X) = \Sigma (x_i - E(X))^2 p_i$$

(parâmetros populacionais)

Uma v.a. continua X é caracterizada por uma função densidade de probabilidade(f.d.p.) f(x), tal que:

a) $f(x) >= 0$, qq x;

b) $\int f(x)\, dx = 1$.

A função f(x) nos diz onde X tem maior "chance" de ocorrer.

Em geral, desconhecemos a distribuição de probabilidade de uma v.a. definida sobre uma população.

Problema Importante:

Formular modelos probabilísticos para variáveis que caracterizam determinadas populações.

Procedimentos:

a) Colher uma amostra da população e definir a v.a. sobre os elementos da amostra, obter informações desejadas a partir dos valores amostrais e tentar inferir conclusões para a população toda. (inferência estatística).

b) se pudemos supor que a dist. de prob. de uma v.a. X possa ser representada por uma dist. Particular → estimar os parâmetros que caracteriza essa particular dist.

Alguns modelos de interesse prático:

a) Bernoulli; Binomial;

b) Poisson; Geométrica; Hipergeométrica;

c) Uniforme; Exponencial;

d) Normal; Gama; quiquadrado; t

Caso discreto: a função de probabilidade $p(x)=P(X=x)$

Caso contínuo: f(x) f.d.p.

- Para gerar as probabilidades de X ~ Bin(12; 0,25) usando Minitab:

> PDF;

> BINOMIAL 12 0.25.

-Para gerar as probabilidades de Y ~ Pois(2,7) usando Minitab

> PDF;

> POISSON 2.7.

No MINITAB,

> pdf;
> bino 12 0,25.

Probability Density Function

Binomial with $n = 12$ and $p = 0,25$

x	P(X = x)
0	0,0317
1	0,1267
2	0,2323
3	0,2581
4	0,1936
5	0,1032
6	0,0401
7	0,0115
8	0,0024
9	0,0004
10	0,0000

Portanto,

$P(X \geq 6) = 0,0544.$

Temos que
$E(X) = n \times p = 12 \text{x} 0,25 = 3,$

Amostra de UMA População

Sejam X_1, ..., X_n: iid com $E(X_i)=E(X)$, $Var(X_i)=Var(X)$

x_1, ..., x_n: os valores observados das v.a. X_1,..., X_n

$x_{(1)}$, ..., $x_{(n)}$: os valores observados ordenados em ordem crescente, são os valores da estatística de ordem $X_{(1)}$, ..., $X_{(n)}$.

Uma análise exploratória de dados(AED) procura responder a seguinte questão:

O que os dados da amostra nos dizem sobre a população da qual eles foram selecionados?

Algumas Questões interessantes:

a) Quais são alguns valores atípicos da amostra?

b) Qual é o valor que representa a posição central do conjunto de dados?

c) Qual é a medida de variabilidade ou dispersão presente nos dados?

d) Os dados podem ser considerados simétricos?

Uma primeira tarifa da AED:

Exibir os dados de maneira apropriada e para isso utilizamos alguns resumos:

a) Dispositivos gráficos e tabelas;

b) Medidas resumo (medidas de posição e dispersão).

Medidas de Posição e Dispersão

MEDIDAS DE POSIÇÃO:

Média, Mediana, Média aparada, Quantis

MEDIDAS DE DISPERSÃO:

Amplitude, Intervalo-Interquartil, Variância, o Desvio médio, Desvio Padrão, Desvio mediano absoluto, variância aparadaCoeficiente de Variação.

- **Média:**

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Dados: 2, 5, 3, 7, 8

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

- **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

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

- **α-média aparada:**

Para α: 0<α<1, α-média aparada,

$$\bar{x}(\alpha)$$

é o valor obtido eliminando-se as $100^{\alpha}\%$ primeiras obs. E as $100^{\alpha}\%$ últimas obs. E tomando a média das obs. restantes.

Ex: obs.: 1,2,3,4,5,6,50,8,9,10

Média=9,8; mediana=5,5

media aparada a 20%=5,8

Ex: obs.: 1,2,3,4,5,6,50,8,9,10

Média=9,8; mediana=5,5

media aparada a 20%=5,8

-A média é bastante afetada pelo valor atípico 50, ao passo que a mediana e a média aparada a 20% não são.

-Essas duas últimas são medidas resistentes.

-Medidas resistentes: muda pouco se alterarmos um número pequeno dos dados.

P-quantil:

É um valor que deixa 100p% das obs. à sua esquerda, com 0<p<1. (definição informal)

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

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

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

$\Rightarrow n=11$

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

- um gráfico de quantis(p x Q(p)) é um auxiliar importante na análise de dados;

- se os valores forem aproximadamente simétricos, a inclinação na parte superior do gráfico deve ser aproximadamente igual à inclinação na parte inferior ;

- os cinco valores (os extremos e os quartis), $x_{(1)}$, Q_1, Q_2, Q_3, $x_{(n)}$, são medidas de localização importantes para avaliarmos a simetria dos dados.

Simetria dos dados:

a) $Q_2 - x_{(1)} \approx x_{(n)} - Q_2$

b) $Q_2 - Q_1 \approx Q_3 - Q_2$

c) $Q_1 - x_{(1)} \approx x_{(n)} - Q_3$

d) $Q_2 - x_{(i)} \approx x_{(n+1-i)} - Q_2$, $i = 1,2, \ldots, [(n+1)/2]$

Assimetria à direita

a) $Q_d - Q_2 \geq Q_2 - Q_e$,

onde Q_d: quantis a direita da mediana,

Q_e: quantis à esquerda da mediana.

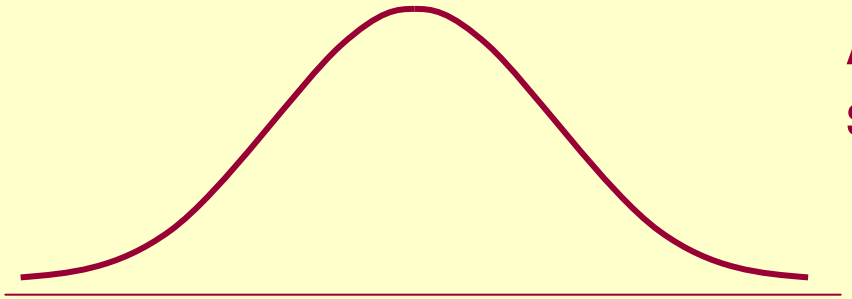
Gráfico de simetria (u_i x v_i)

$$u_i = Q_2 - x_{(i)}$$

$$v_i = x_{(n+1-i)} - Q_2$$

Se a distribuição dos dados for simétrica, os pontos (u_i, v_i) deverão estar ao longo da reta u=v.

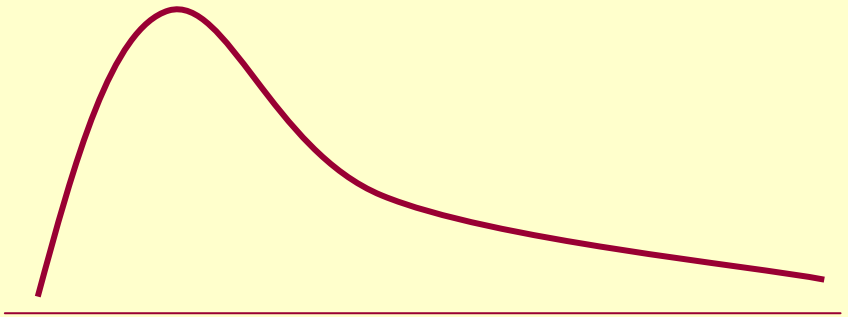
Simetria



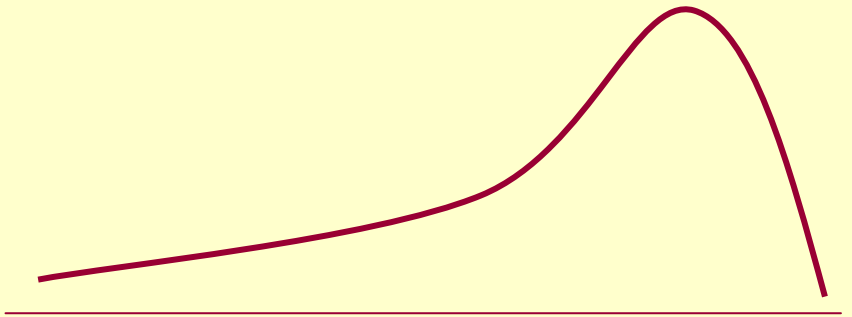
Distribuição Simétrica

As variáveis (Y-c) e –(Y-c) são identicamente distribuídas

$$F_{Y-c}(\cdot) = F_{-(Y-c)}(\cdot)$$



Assimetria à Direita



Assimetria à Esquerda

Coeficientes de Assimetria

- **Coeficiente de Bowley**

$$Ass_B = \frac{(Q_3 - Me) - (Me - Q_1)}{Q_3 - Q_1}$$

- **Coeficiente de Pearson**

$$Ass_P = \frac{3\left(\overline{Y} - Me\right)}{s}$$

|Ass| < 0,15 $\Rightarrow$ Grau de Assimetria pequeno

|Ass| > 1 $\Rightarrow$ Grau de Assimetria elevado

0,15 < |Ass| < 1 $\Rightarrow$ Grau de Assimetria moderado

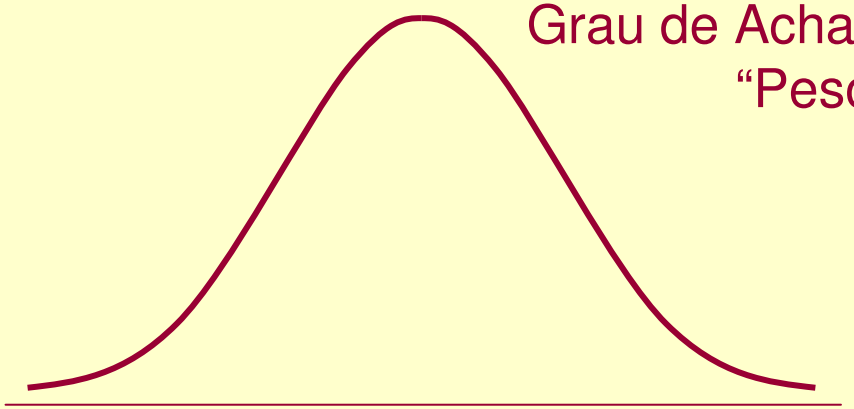
- **Coeficiente de Fisher**

$$Ass_F = \frac{m_3}{s^3} = \frac{\left(\sum (Y_j - \overline{Y}\right)^3 / n}{s^3}$$

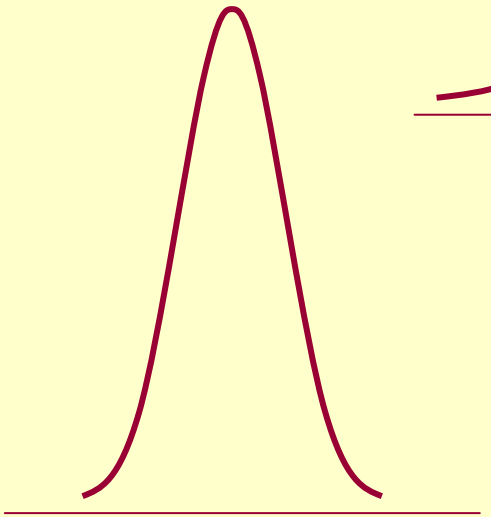
Qual é o valor destes coeficientes para a Normal ?

Curtose

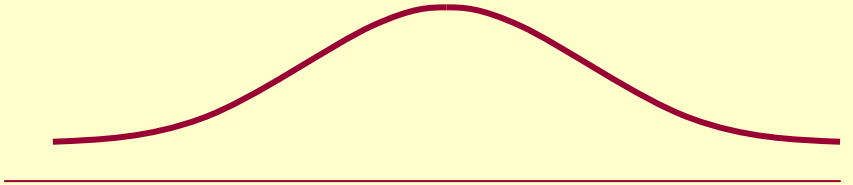
Grau de Achatamento da Distribuição
"Peso das Caudas"



Distribuição Mesocúrtica



Distribuição Leptocúrtica
"Cauda Leve"



Distribuição Platicúrtica
"Cauda Pesada"

Coeficiente de Curtose

- **Excesso de Curtose**

$$ExCur = \frac{m_4}{s^4} - 3$$

$$m_4 = \frac{\sum (Y_j - \bar{Y})^4}{n}$$

ExCur = 0 $\Rightarrow$ Distribuição Mesocúrtica $\Rightarrow$ Normal

ExCur > 0 $\Rightarrow$ Distribuição Platicúrtica

ExCur < 0 $\Rightarrow$ Distribuição Leptocúrtica

- **Percentílico de Curtose**

$$C = \frac{(Q_3 - Q_1)}{2(P_{90} - P_{10})}$$

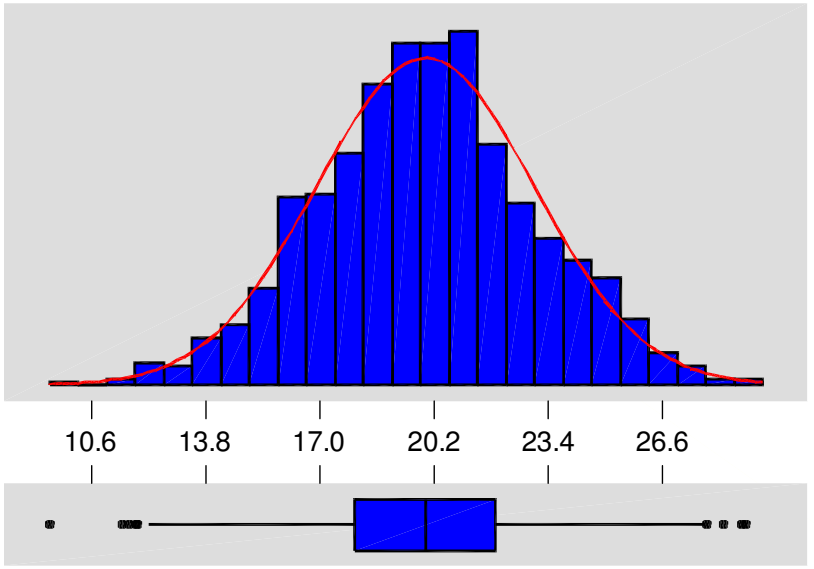
C = 0,263 $\Rightarrow$ Distribuição Mesocúrtica $\Rightarrow$ Normal

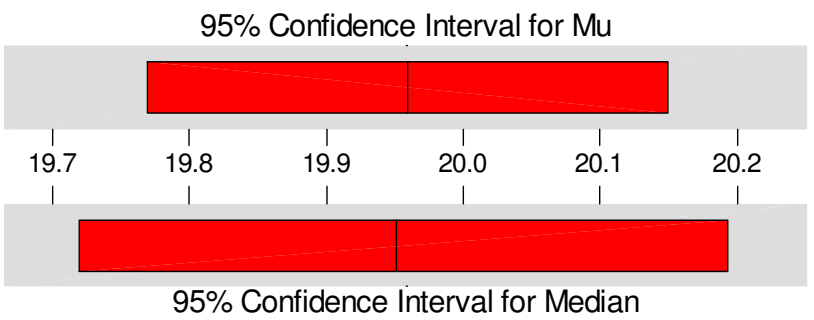
C > 0,263 $\Rightarrow$ Distribuição Platicúrtica

C < 0,263 $\Rightarrow$ Distribuição Leptocúrtica

Calcule o valor de C para a N(0;1) !

Descriptive Statistics - Simulação de 1000 N(20;9)





Variable: Normal

Anderson-Darling Normality Test

A-Squared:	0.564
P-Value:	0.144
Mean	19.9592
StDev	3.0565
Variance	9.34235
Skewness	1.31E-02
Kurtosis	-1.3E-02
N	1000
Minimum	9.4222
1st Quartile	17.9671
Median	19.9510
3rd Quartile	21.9066
Maximum	28.9159

95% Confidence Interval for Mu

19.7695	20.1489

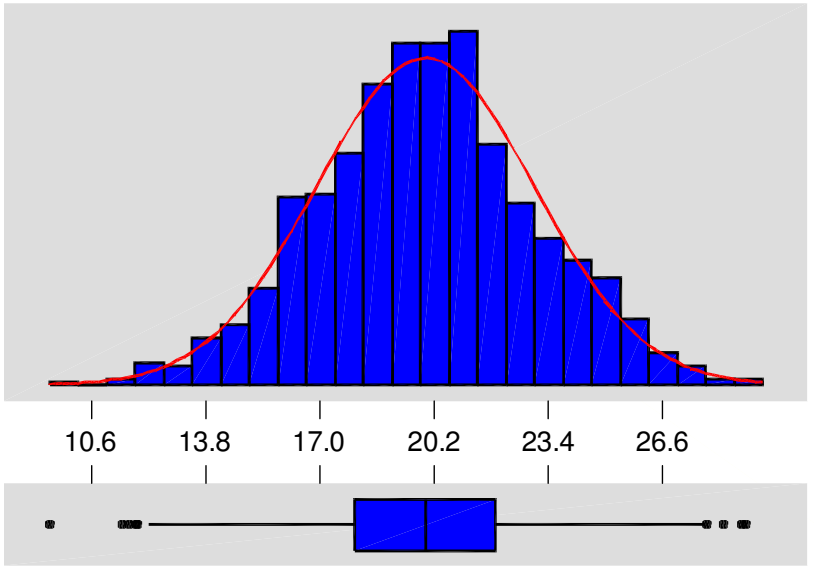
95% Confidence Interval for Sigma

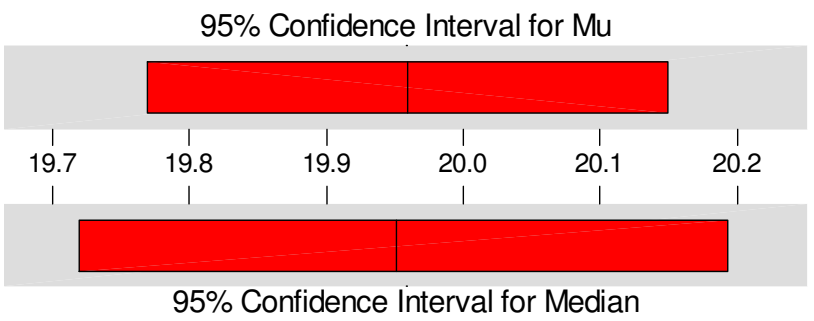
2.9282	3.1967

95% Confidence Interval for Median

19.7188	20.1923

Descriptive Statistics - Simulação de 1000 N(20;9)





Variable: Normal

Anderson-Darling Normality Test

A-Squared:	0.564
P-Value:	0.144
Mean	19.9592
StDev	3.0565
Variance	9.34235
Skewness	1.31E-02
Kurtosis	-1.3E-02
N	1000
Minimum	9.4222
1st Quartile	17.9671
Median	19.9510
3rd Quartile	21.9066
Maximum	28.9159

95% Confidence Interval for Mu

19.7695	20.1489

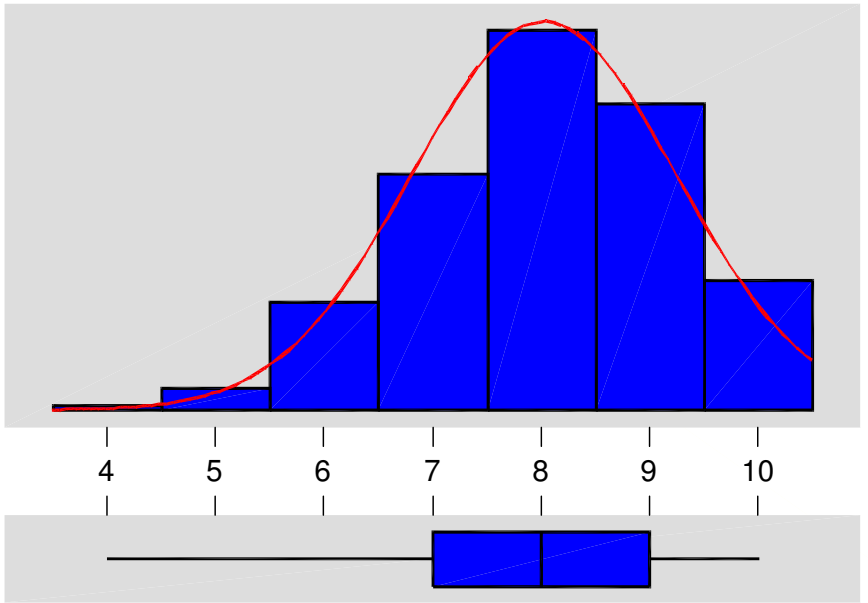
95% Confidence Interval for Sigma

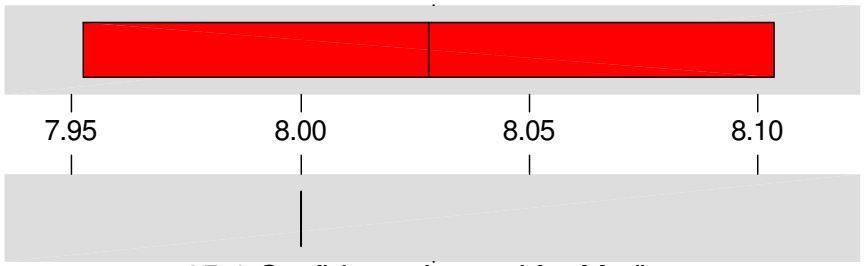
2.9282	3.1967

95% Confidence Interval for Median

19.7188	20.1923

Descriptive Statistics - Simulação de 1000 Bino(10;0.8)





Variable: Bino

Anderson-Darling Normality Test

A-Squared:	29.313
P-Value:	0.000

Mean	8.02800
StDev	1.21684
Variance	1.48070
Skewness	-3.6E-01
Kurtosis	-1.7E-01
N	1000

Minimum	4.0000
1st Quartile	7.0000
Median	8.0000
3rd Quartile	9.0000
Maximum	10.0000

95% Confidence Interval for Mu

7.9525	8.1035

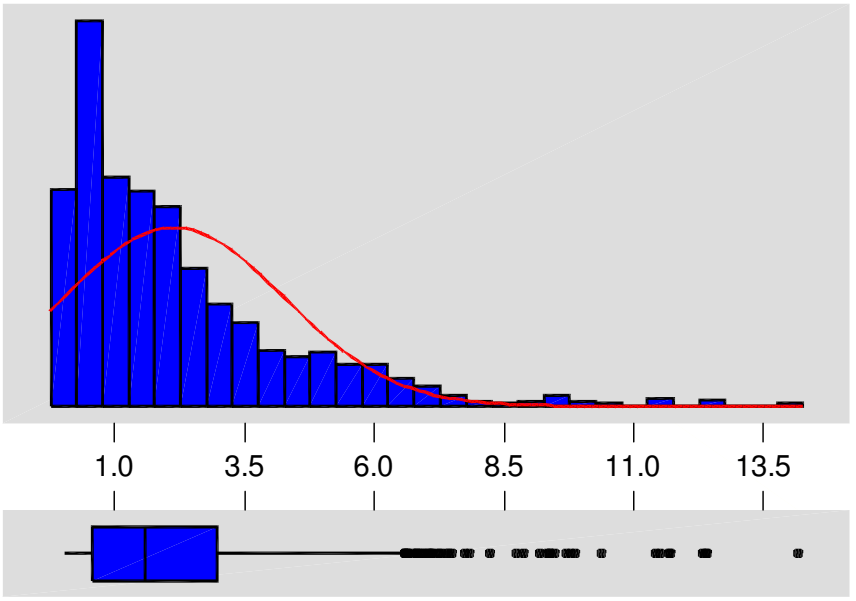
95% Confidence Interval for Sigma

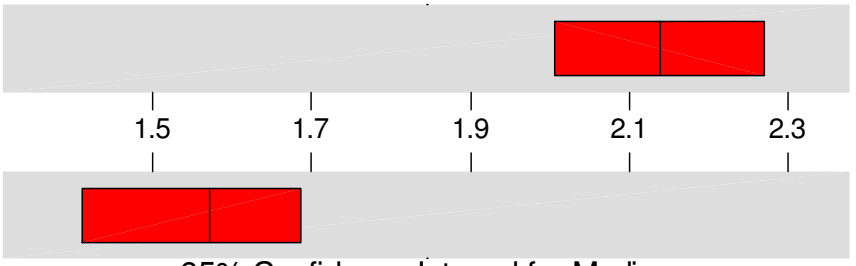
1.1657	1.2727

95% Confidence Interval for Median

8.0000	8.0000

Descriptive Statistics - Simulação de 1000 Chi(2)





Variable: Chi

Anderson-Darling Normality Test

A-Squared:	45.388
P-Value:	0.000
Mean	2.13845
StDev	2.12922
Variance	4.53357
Skewness	1.81217
Kurtosis	4.30253
N	1000
Minimum	0.0024
1st Quartile	0.5541
Median	1.5707
3rd Quartile	2.9603
Maximum	14.1631

95% Confidence Interval for Mu

2.0063	2.2706

95% Confidence Interval for Sigma

2.0398	2.2269

95% Confidence Interval for Median

1.4091	1.6853