ESTATÍSTICA

Aplicada

Prof. Dr. Ricardo da Silva Manca

Recordando... Tipos de Estudos

- **Estudos Observacionais**: Observamos e medimos características específicas, mas não tentamos modificar os sujeitos objeto do estudo.
- Ex. Entrevistas no geral, pesquisa de ibope.
- **Experimento**: Aplicamos algum tratamento e passamos, então, a observar seu efeito sobre os sujeitos (unidades experimentais.
- Ex. Teste de um novo medicamento.

Tipos de Amostras

- **Amostra Aleatória**: Neste caso, membros de uma população são selecionados de tal modo que cada MEMBRO INDIVIDUAL tenha chance IGUAL de ser selecionado.
- **Amostra Aleatória Simples**: Uma amostra aleatória simples de tamanho *n* é selecionada de tal modo que toda AMOSTRA POSSÍVEL DE MESMO TAMANHO *n* tenha a mesma chance de ser escolhida.
- **Amostra Probabilística**: Envolve a seleção de membros de uma população de tal modo que cada membro tenha uma chance conhecida (NÃO NECESSARIAMENTE IGUAL) de ser selecionado.

Amostragens

- **Amostragem Aleatória**: Cada membro da população tem chance igual de ser escolhido.

Amostragens

- **Amostragem Sistemática**: Escolha algum ponto inicial e a seguir selecione cada *k*-ésimo elemento da população.

Amostragens

- **Amostragem de Conveniência**: Usa resultados que são fáceis de coletar.

Amostragens

- **Amostragem Estratificada**: Subdivida a população em pelo menos dois subgrupos diferentes (ou estratos), de modo que os sujeitos dentro do mesmo subgrupo tenham as mesmas características (ex. Sexo ou faixa etária) e a seguir extraia uma amostra de cada subgrupo.

Amostragens

- **Amostragem por conglomerado**: Divida a população em seções (ou conglomerados), a seguir selecione aleatoriamente alguns desses conglomerados e escolha todos os membros desses conglomerados selecionados.

Erros Amostrais

- Mesmo que o processo de coleta da amostra seja bem planejado, provavelmente sempre haverá algum erro nos resultados.
- **Erro Amostral**: É a diferença entre o resultado amostral e o verdadeiro resultado da população; tais erros resultam das flutuações amostrais devido ao acaso.
- **Erro não-amostral**: Ocorre quando os dados amostrais são coletados, registrados ou analisados incorretamente (como a seleção de uma amostra tendenciosa, uso de instrumento defeituoso, etc.)

Características Importantes dos Dados

- **Centro**: Um valor representativo ou médio, que indica onde se localiza o meio do conjunto de dados.
- **Variação**: Uma medida de quanto os valores dos dados variam entre eles.
- **Distribuição**: A natureza ou forma da distribuição dos dados (tal como em forma de sino, uniforme ou assimétrica).
- ***Outliers* ou Valores Discrepantes**: Valores amostrais que se localizam muito longe da grande maioria dos outros valores amostrais.
- **Tempo**: Características dos dados que mudam com o tempo.

Distribuições de Frequência

Distribuições de Frequência

- Uma **distribuição de frequência** lista os valores dos dados (individualmente ou por grupos de intervalos), juntamente com suas frequências correspondentes (ou contagens).

Distribuições de Frequência

TABELA Notas Obtidas por 500 Alunos em um Teste de Estatística

Notas	f_j	fr_j (%)	F_j	De acordo com o item (a)	De acordo com o item (b)
				Fr_j (%)	Fr_j
0 ⊢ 10	5	1	5	1	5/500 = 0,01 ou 1%
10 ⊢ 20	15	3	20	4	20/500 = 0,04 ou 4%
20 ⊢ 30	20	4	40	8	40/500 = 0,08 ou 8%
30 ⊢ 40	45	9	85	17	85/500 = 0,17 ou 17%
40 ⊢ 50	100	20	185	37	185/500 = 0,37 ou 37%
50 ⊢ 60	130	26	315	63	315/500 = 0,63 ou 63%
60 ⊢ 70	100	20	415	83	415/500 = 0,83 ou 83%
70 ⊢ 80	60	12	475	95	475/500 = 0,95 ou 95%
80 ⊢ 90	15	3	490	98	490/500 = 0,98 ou 98%
90 ⊢ 100	10	2	500	100	500/500 = 1,00 ou 100%
	500	100			

Distribuições de Frequência

- **Rol**: é uma lista em que os valores estão dispostos em uma determinada ordem, crescente ou decrescente;
- **Limites inferiores de classe**: são os menores números que podem pertencer às diferentes classes;
- **Limites superiores de classe**: são os maiores números que podem pertencer às diferentes classes;
- **Pontos médios de classe**: são os pontos médios dos intervalos que determinam cada classe.
- **Amplitude de classe**: é a diferença entre dois limites de classe consecutivos.

Distribuições de Frequência

As distribuições de frequência são construídas pelas seguintes razões:

- Grandes conjuntos de dados podem ser resumidos;
- Podemos obter alguma compreensão sobre a natureza dos dados;
- Temos uma base para construir gráficos importantes (tal como o histograma).

Distribuições de Frequência

- **Frequências relativas**: Divide-se cada frequência de classe pelo total de todas as frequências.

$$\text{Frequência Relativa} = \frac{\text{Frequência de Classe}}{\text{Soma de todas as frequências}}$$

- **Frequência acumulada**: A frequencia acumulada de uma classe é a soma da frequência daquela classe mais as frequências de todas as classes anteriores.

Elaboração de uma distribuição de frequência

1. Liste os dados brutos que podem ou não serem transformados em um rol.
2. Encontre a amplitude total A_t do conjunto de valores observados (A_t = Valor máximo – Valor mínimo)
3. Defina o número de classes a serem utilizadas.

Como sugestão, pode-se utilizar o **Critério de Sturges**:

nº de classes = $1 + 3{,}3 \times \log_{10} n$, onde n = número de observações.

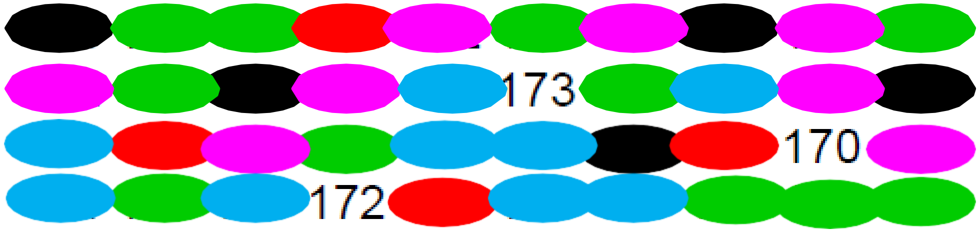
4. Determine a amplitude do intervalo de classe.

A amplitude do intervalo de classe será igual ao quociente entre a amplitude total da série e o número de classes escolhido:

$$\text{Amplitude de Classe} \approx \frac{(\text{Valor Máximo}) - (\text{Valor Mínimo})}{\text{Número de Classes}}$$

Distribuições de Frequência

- **<u>EXEMPLO</u>:**
- Considere a relação de número abaixo, referente às alturas (em centímetros) dos alunos da UTFPR:



- Utilizando o critério de Sturges, tem-se:

número de classes = $1 + 3{,}3 \times \log 40 \approx 6{,}286798 \approx 7$

(**arredonde para cima**)

Distribuições de Frequência

Intervalos de Classe	Ponto médio	Frequência absoluta	Frequência Relativa	Freq. Abs. Acumulada	Freq. Rel. Acumulada
150 ├── 154	(150+154)/2 = 152	4	4 / 40 = 0,100	4	0,100
154 ├── 158	156	9	0,225	4 + 9 = 13	0,325
158 ├── 162	160	11	0,275	24	0,600
162 ├── 166	164	8	0,200	32	0,800
166 ├── 170	168	5	0,125	37	0,925
170 ├── 174	172	3	0,075	40	1,000
		40	1		

Diagramas de Ramo e Folhas

- Representam dados separando cada valor em duas partes:
  - Ramo: Dígito mais à esquerda;
  - Folha: Dígito mais à direita.
- As folhas são arranjadas em ordem crescente, não na ordem em que aparecem na lista original.
- Uma grande vantagem é que podemos ver a distribuição dos dados e ainda reter toda a informação da lista original;
- É uma maneira rápida e fácil de ordenar os dados (Será importante para se encontrar mediana ou percentis).

Exemplo

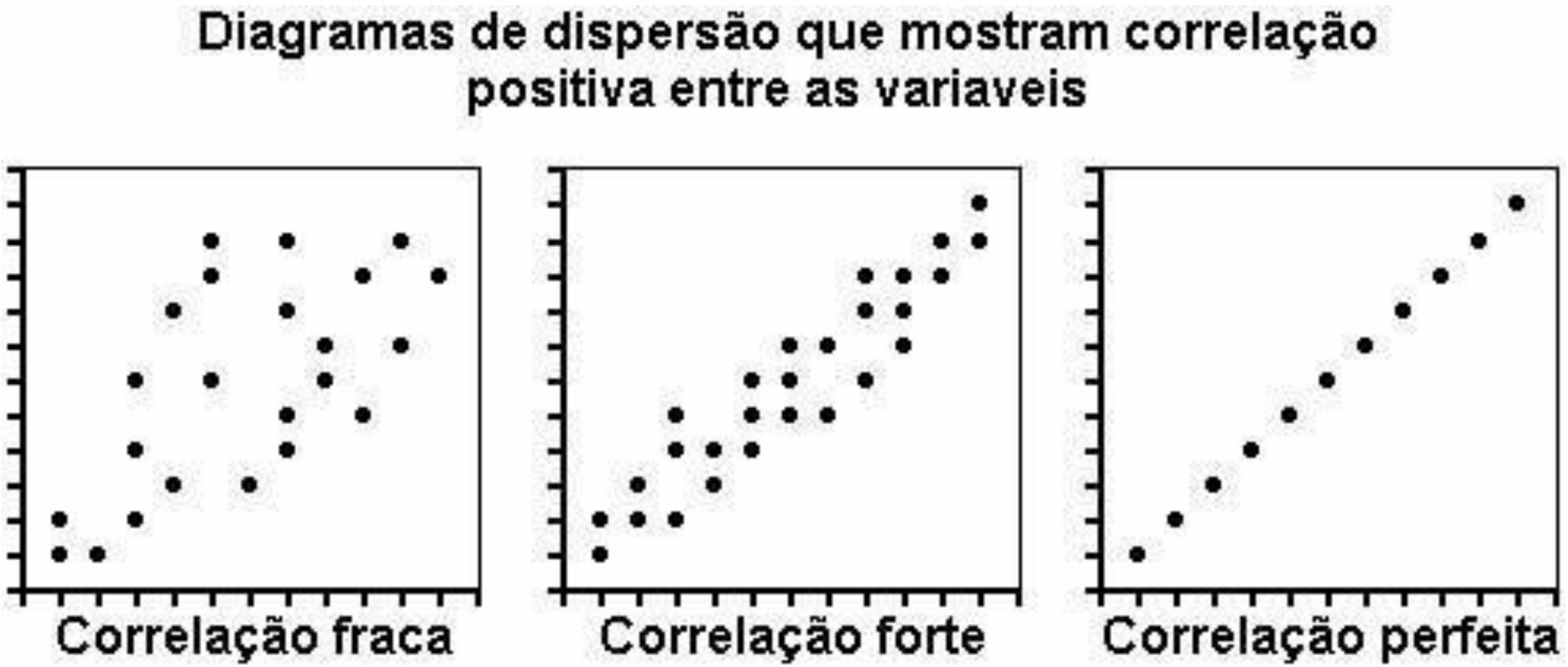
- Altura (em cm) dos alunos da FMPFM.

RAMOS	FOLHAS
15	0 1 2 3 4 5 5 5 5 6 6 6 7 8 8
16	0 0 0 0 0 1 1 1 1 2 2 3 3 4 4 4 5 6 7 8 8 9
17	0 2 3

Diagrama de Dispersão

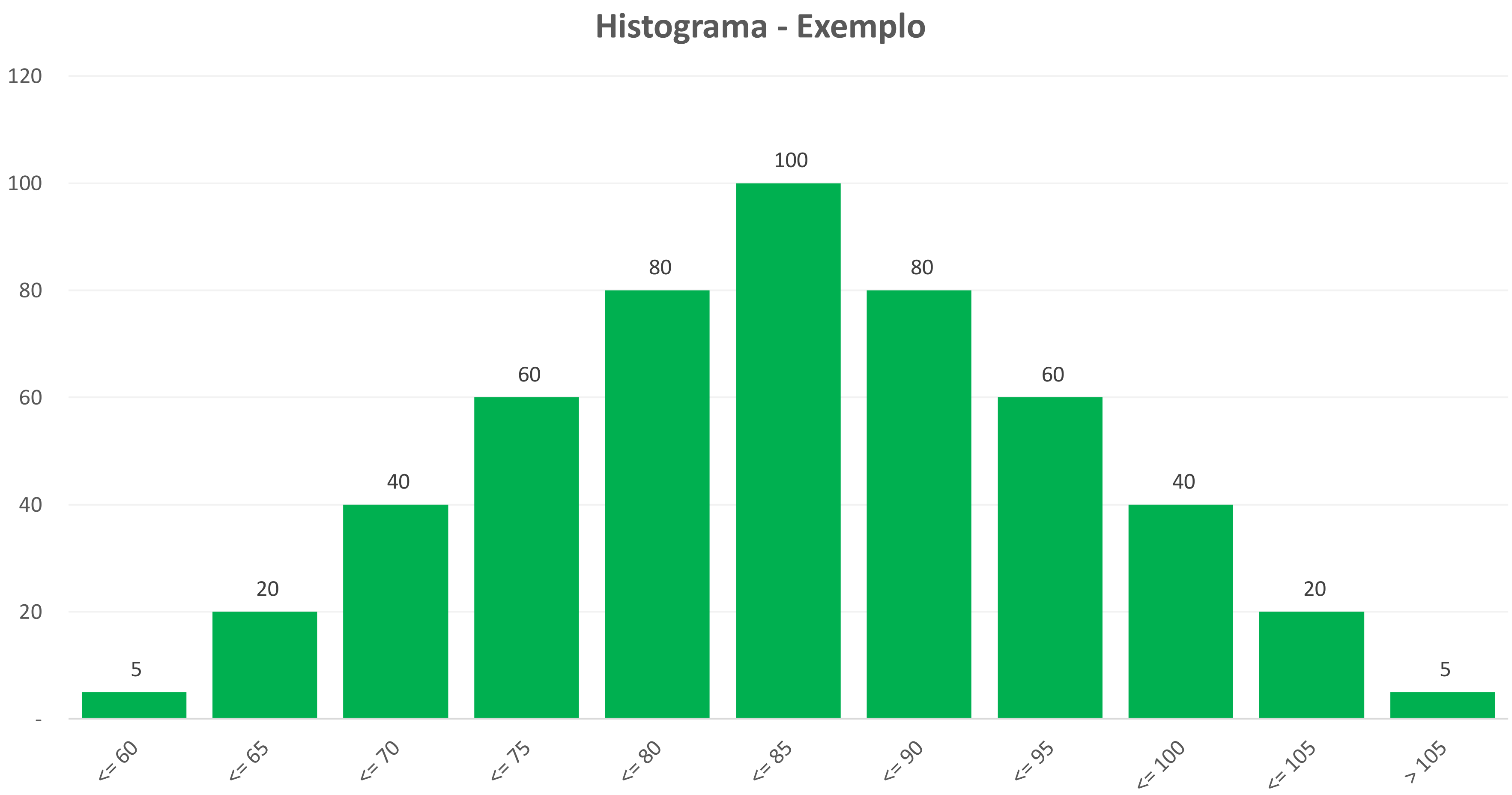
- É um gráfico de pares de dados (x, y), com um eixo x horizontal e um eixo y vertical.
- Os dados são colocados em pares que combinam cada valor de um conjunto de dados com um valor correspondente de um segundo conjunto de dados.
- É útil para se determinar a existência, ou não, de alguma relação entre as variáveis.

Diagrama de Dispersão

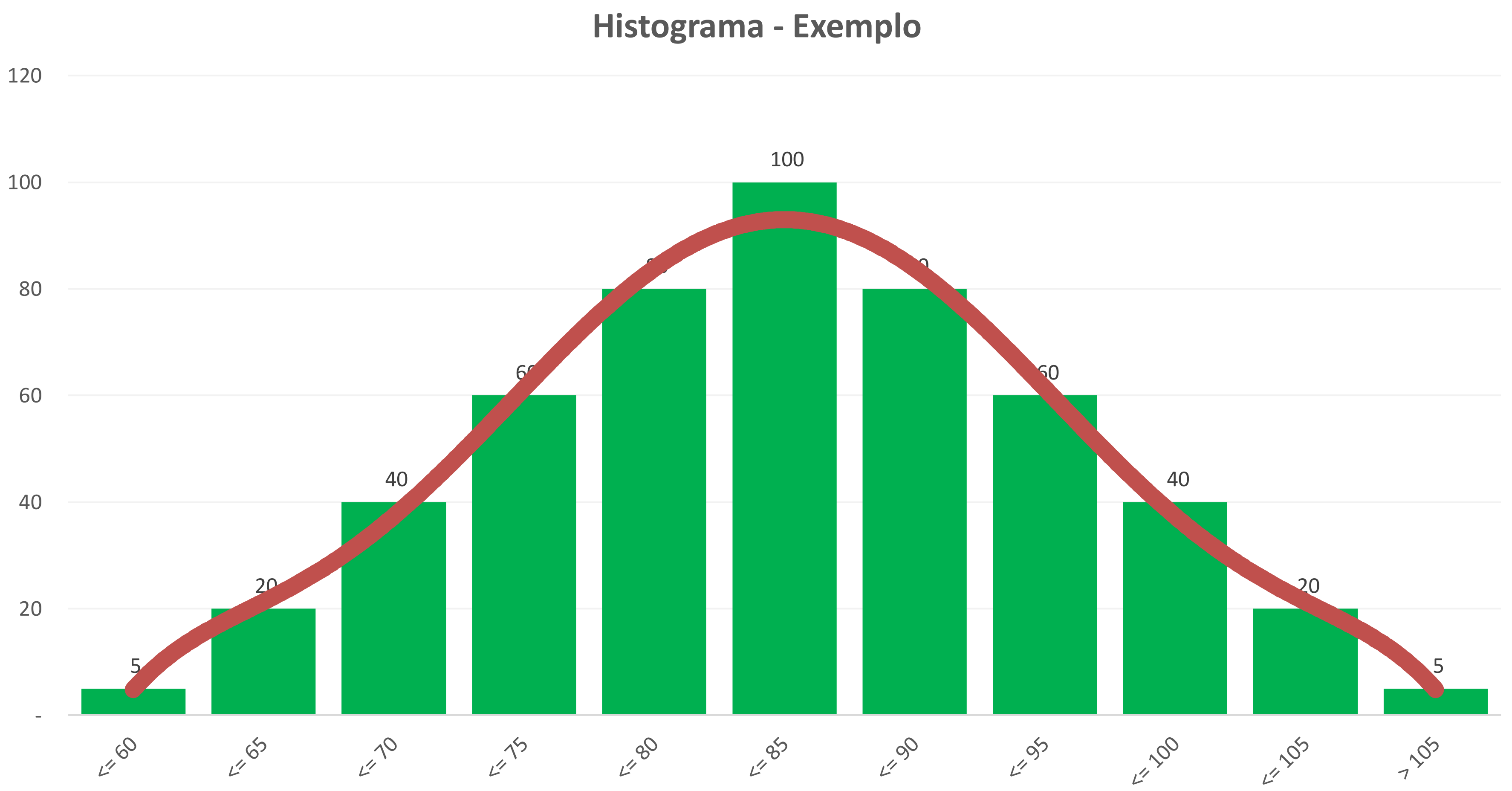


Estudando o Histograma

Histograma - Interpretação



Histograma - Interpretação



Histogramas

- Um histograma é um **gráfico de barras** no qual a escala horizontal representa classes de valores de dados e a escala vertical representa frequências.
- As alturas das barras correspondem aos valores das frequências, e as barras são desenhadas adjacentes umas às outras (sem separação).
- **VERSÃO GRÁFICA** de uma tabela de distribuições de frequência.

Distribuição Normal

- Uma característica essencial de uma distribuição normal é que, quando se constrói seu gráfico o resultado tem forma de "SINO".
- As frequências começam baixas, **crescem** até uma **frequência máxima** e depois **decrescem** para uma frequência baixa.
- A distribuição deve ser aproximadamente SIMÉTRICA, com frequências igualmente distribuídas em ambos os lados da frequência máxima.

Distribuições

- **Distribuição Assimétrica**: Quando se estende mais para um lado do que para o outro.

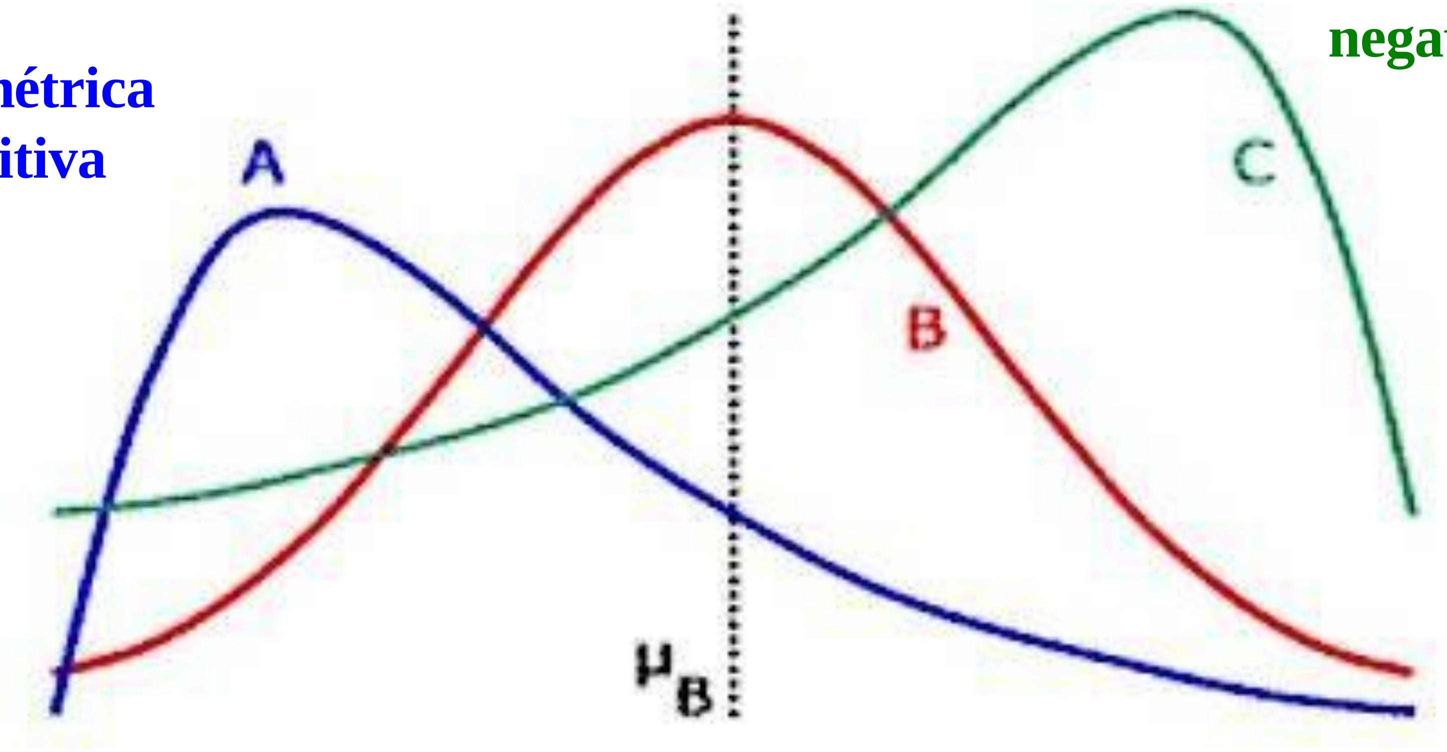
- **Distribuição Simétrica**: Quando a metade esquerda de seu histograma é praticamente uma imagem espelhada de sua metade direita.

Distribuições

Simétrica

Assimétrica negativa

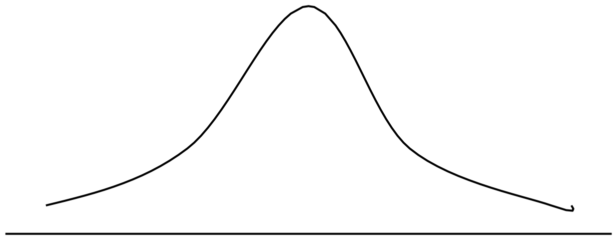
Assimétrica positiva



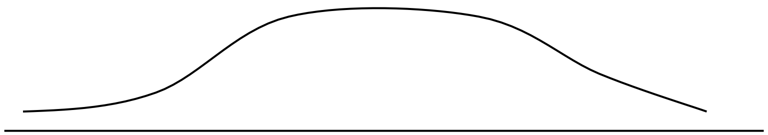
Curtose

- Denomina-se curtose o grau de achatamento da distribuição.

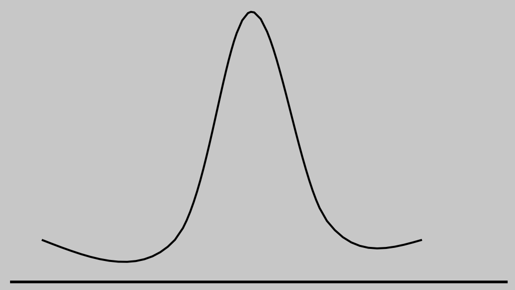
- Uma destituição nem chata e nem delgada, é denominada de mesocúrtica.



- Uma distribuição achatada denomina-se platicúrtica.



- Uma distribuição delgada é denominada de leptocúrtica.



- Para medir o grau de curtose utiliza-se o coeficiente:

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

■Se K = 0,263, diz-se que a curva correspondente à distribuição de freqüência é **mesocúrtica**.

■ Se K > 0,263, diz-se que a curva correspondente à distribuição de freqüência é **platicúrtica.**

■ Se K < 0,263, diz-se que a curva correspondente à distribuição de freqüência é **leptocúrtica.**

Estudando a Distribuição Normal

Distribuição Normal

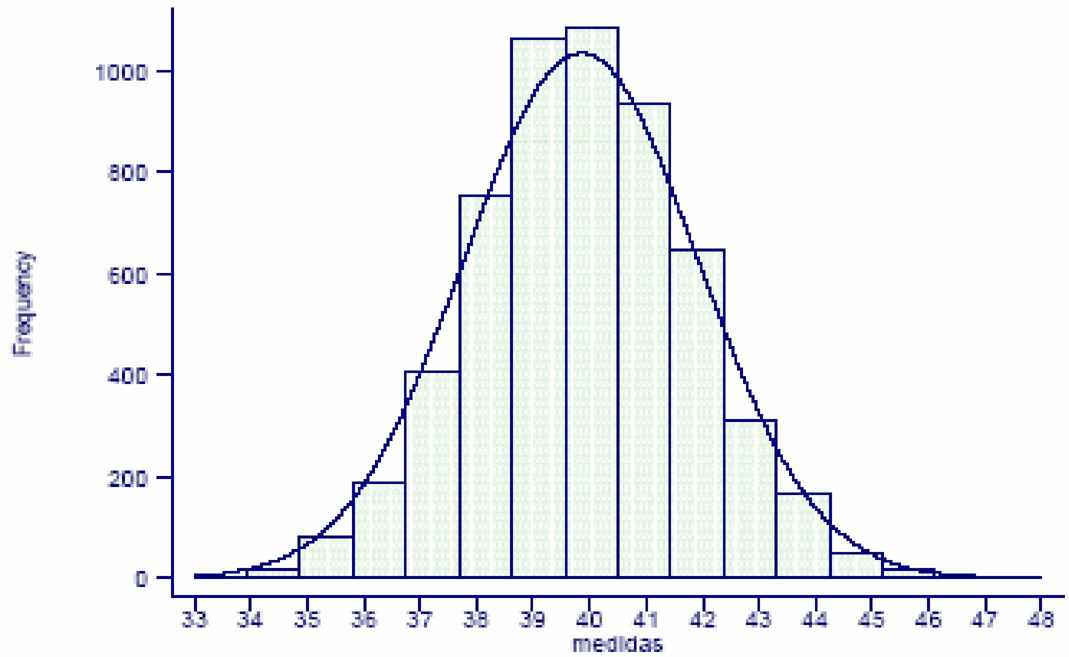
- Muitas variáveis estudadas na área biomédica apresentam distribuição simétrica (os valores centrais são mais freqüentes e os valores extremos mais raros).

- Na prática, se o coeficiente de assimetria está situado no intervalo (-0.5,+0.5), considera-se a distribuição aproximadamente simétrica.

- Uma distribuição simétrica típica é a distribuição normal.

Exemplo: Distribuição Normal

Distribuição de medidas do tórax (polegadas) de soldados escoceses



Fonte: Daly F et al. Elements of Statistics, 1999

Distribuição Normal

- Por que é importante que as variáveis possam ser descritas por uma distribuição normal?

- Motivo é simples: Se as variáveis respeitam uma distribuição normal, pode-se aplicar a grande maioria dos testes e métodos estatísticos conhecidos.

  ➔ Tem-se maior facilidade

- Variáveis que não têm distribuição normal podem ser submetidas a transformações (raiz quadrada, logaritmo)

Propriedades da Distribuição Normal

- A distribuição é simétrica: Média = mediana = moda.

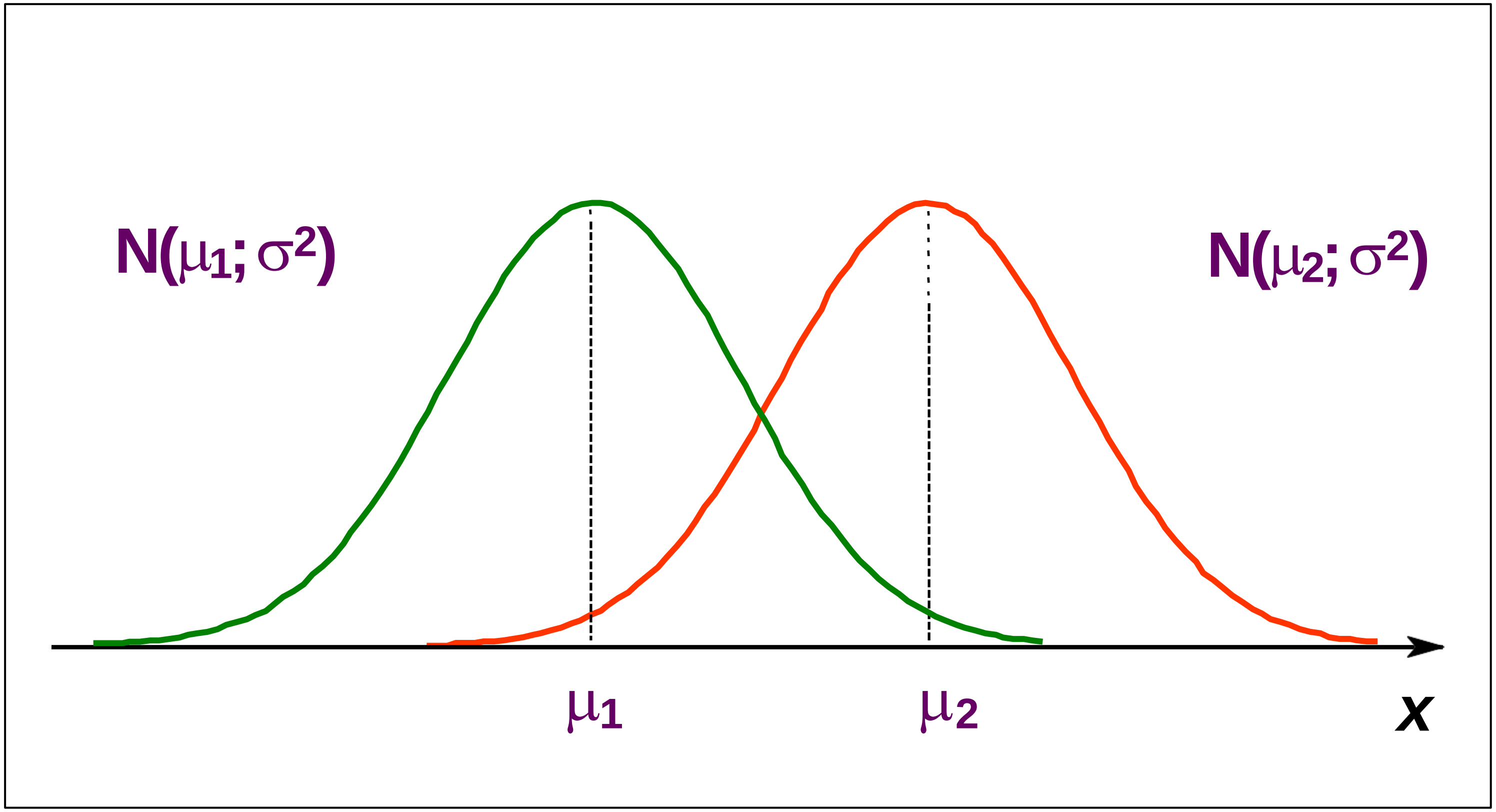
- Os parâmetros μ (média) e σ (desvio padrão) definem completamente uma curva normal. Notação: $X \sim N(\mu,\sigma^2)$

- Na distribuição normal com média **μ** e desvio padrão **σ**:

  - 68% das observações estão a menos de $\pm\sigma$ da média μ.
  - 95% das observações estão a menos de $\pm 2\sigma$ de μ.
  - 99.5% das observações estão a menos de $\pm 3\sigma$ de μ.

Propriedades da Distribuição Normal

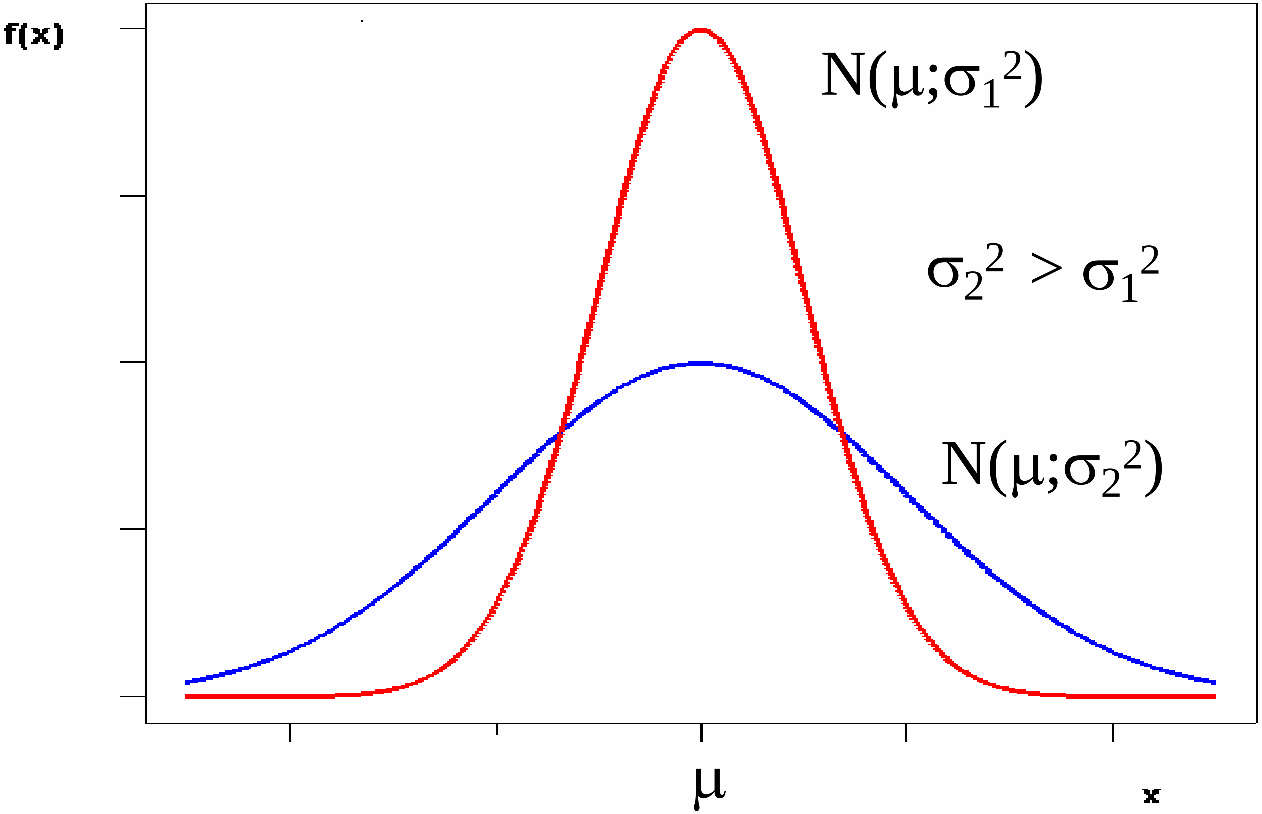
A distribuição Normal depende dos parâmetros μ e σ^2



Curvas Normais com mesma variância σ^2
mas médias diferentes ($\mu_2 > \mu_1$).

Propriedades da Distribuição Normal

Influência de σ^2 na curva Normal



Curvas Normais com mesma média μ,
mas com variâncias diferentes ($\sigma_2^2 > \sigma_1^2$).

Distribuição Normal

- A distribuição normal pode ser descrita pela seguinte "função de densidade":

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \times \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \quad , -\infty < x < +\infty$$

- A área total embaixo da curva normal é igual a 1.

- Quando temos em mãos uma variável aleatória com distribuição normal, nosso principal interesse é obter a probabilidade dessa variável aleatória assumir um valor em um determinado intervalo.

Distribuição Normal Padrão

- Caso especial da distribuição Normal: N(0,1).

- Para transformar uma variável de forma que tenha média 0 e desvio padrão 1 (padronização ou normalização), basta fazer o cálculo:

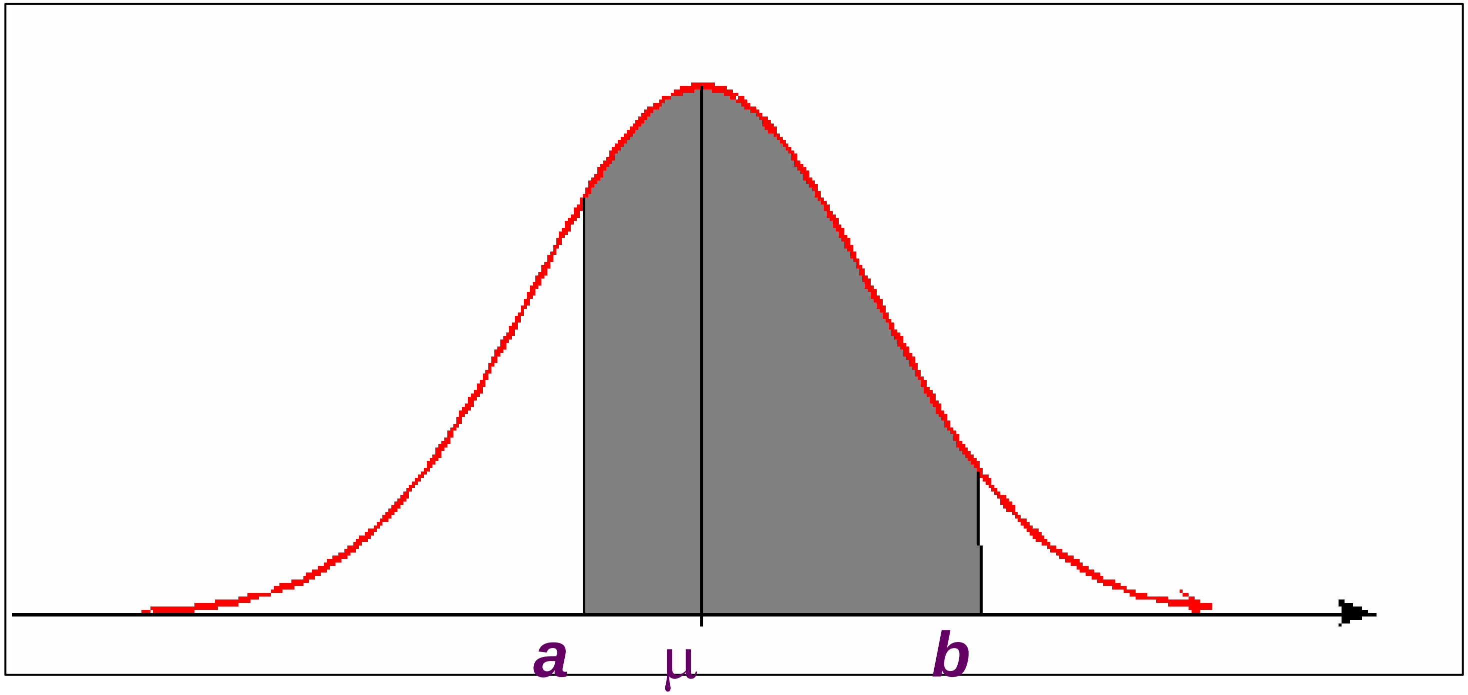
$$\boxed{Z = \frac{X - \mu}{\sigma}}$$

- Propriedade dessa distribuição: Podemos calcular probabilidades usando a tabela da distribuição normal padronizada.

Cálculo de probabilidades

P(*a* < *X* < *b*)

Área sob a curva e acima do eixo horizontal (*x*) entre *a* e *b*.



Usando escores Z para determinar probabilidades:

Se $X \sim N(\mu \,;\, \sigma^2)$, definimos $\boxed{Z = \dfrac{X - \mu}{\sigma}}$

Portanto,

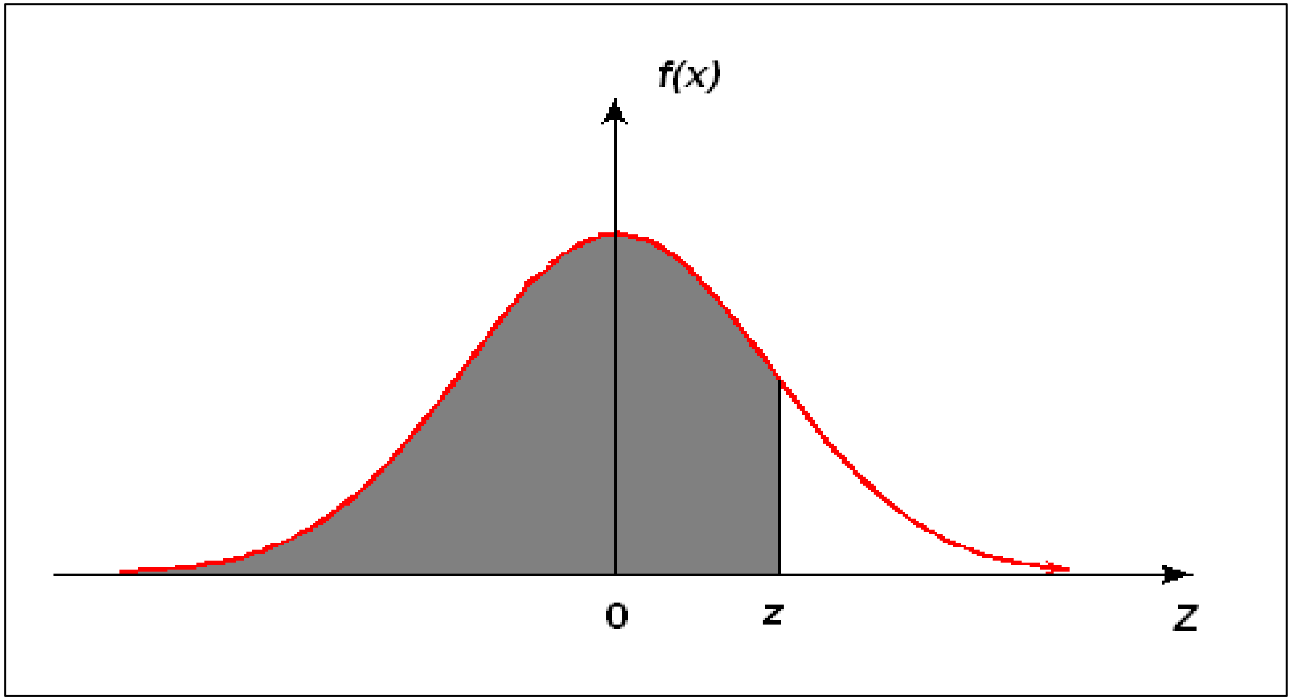
$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

Exemplo: Seja $X \sim N(10 \,;\, 64)$ ($\mu = 10$, $\sigma^2 = 64$ e $\sigma = 8$).
Calcular $P(6 \le X \le 12)$.

$$P(6 \le X \le 12) = P\left(\frac{6-10}{8} < \frac{X-10}{8} < \frac{12-10}{8}\right) = P(-0{,}5 < Z < 0{,}25)$$

Para cálculo dessa probabilidade utilizamos a tabela normal padrão.

USO DA TABELA NORMAL PADRÃO



Denotamos : A(z) = P($Z \leq z$), para $z \geq 0$.

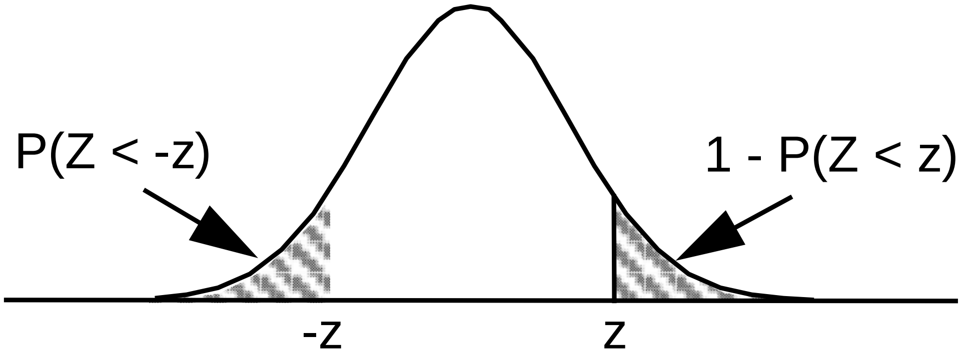
USO DA TABELA NORMAL PADRÃO

As propriedades que seguem podem ser deduzidas da simetria da densidade em relação à média 0, e são úteis na obtenção de outras áreas não tabuladas.

1. $P(Z>z) = 1 - P(Z<z)$

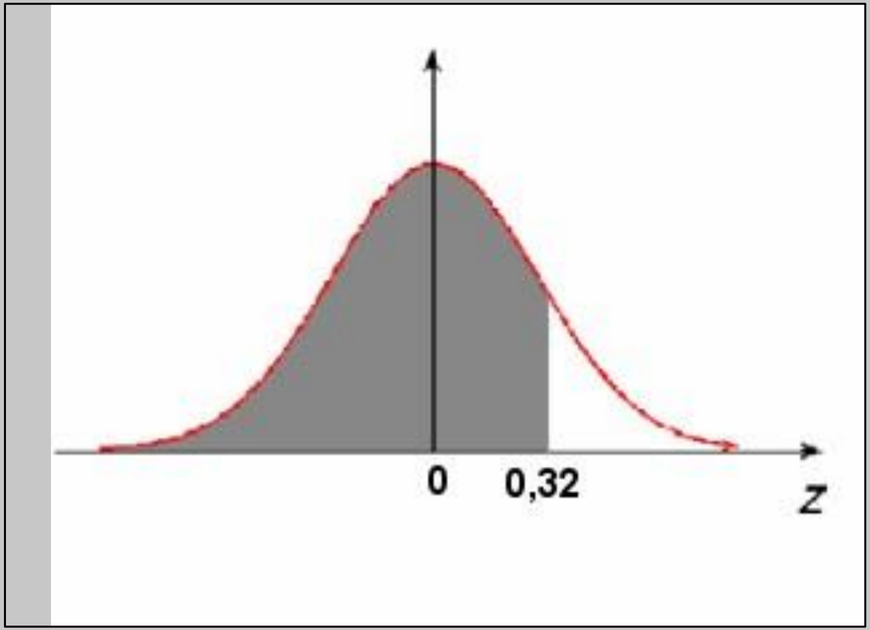
2. $P(Z<-z) = P(Z>z)$

3. $P(Z>-z) = P(Z<z)$.



Exemplo: Seja $Z \sim N(0; 1)$, calcular

a) $P(Z \leq 0{,}32)$

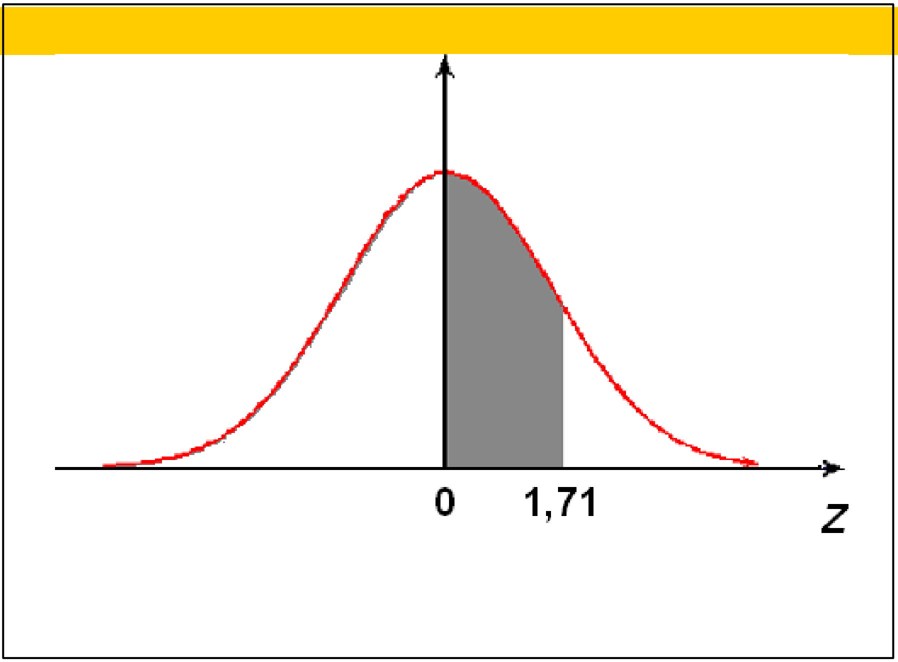


$P(Z \leq 0{,}32) = A(0{,}32) = 0{,}6255.$

Encontrando o valor na Tabela N(0;1):

z	0	1	2
0,0	0,5000	0,5039	0,5079
0,1	0,5398	0,5437	0,5477
0,2	0,5792	0,5831	0,5870
0,3	0,6179	0,6217	0,6255

b) $P(0 < Z \leq 1{,}71)$



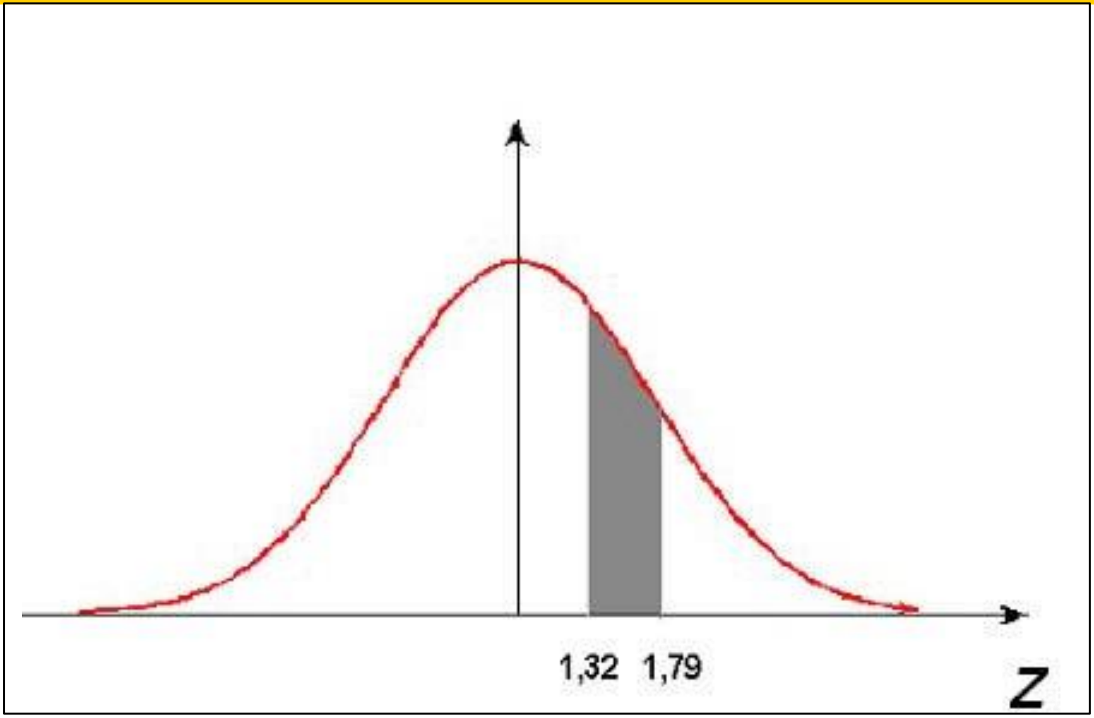
$$P(0 < Z \leq 1{,}71) = P(Z \leq 1{,}71) - P(Z \leq 0)$$

$$= A(1{,}71) - A(0)$$

$$= 0{,}9564 - 0{,}5 = 0{,}4564.$$

Obs.: $P(Z < 0) = P(Z > 0) = 0{,}5.$

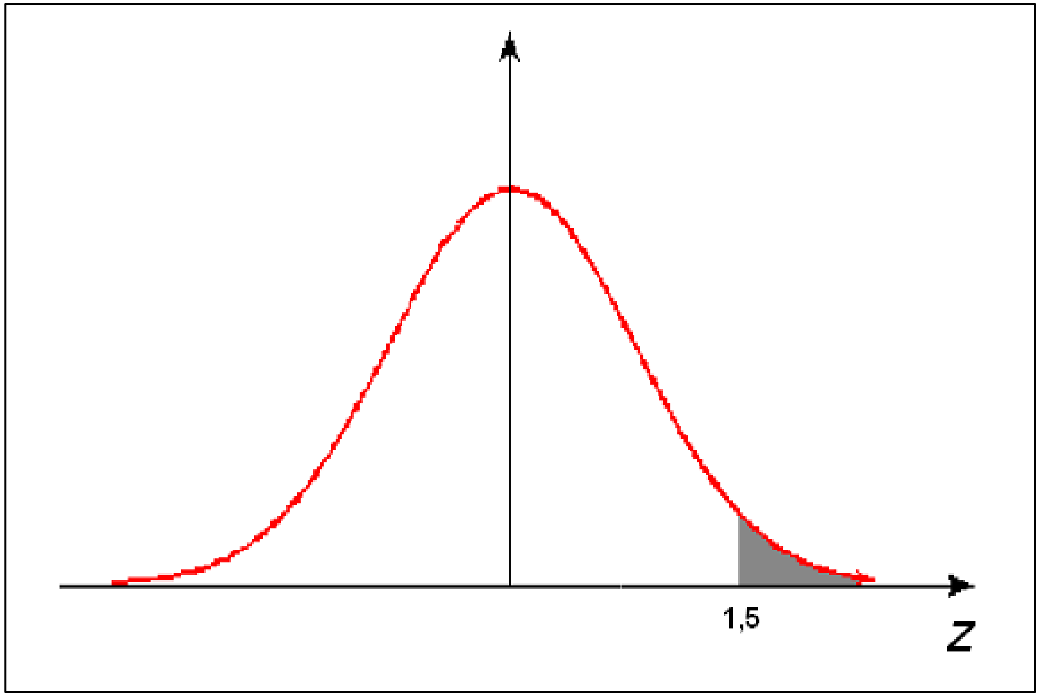
c) $P(1{,}32 < Z \leq 1{,}79)$



$$P(1{,}32 < Z \leq 1{,}79) = P(Z \leq 1{,}79) - P(Z \leq 1{,}32) = A(1{,}79) - A(1{,}32)$$

$$= 0{,}9633 - 0{,}9066 = 0{,}0567.$$

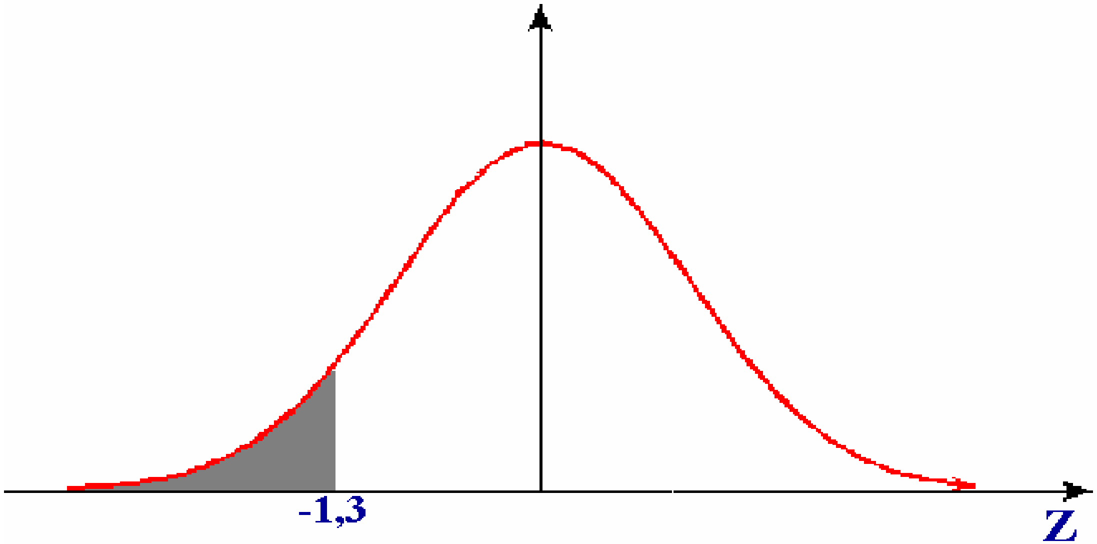
d) $P(Z \geq 1,5)$



$$P(Z > 1,5) = 1 - P(Z \leq 1,5) = 1 - A(1,5)$$

$$= 1 - 0,9332 = 0,0668.$$

e) $P(Z \leq -1,3)$

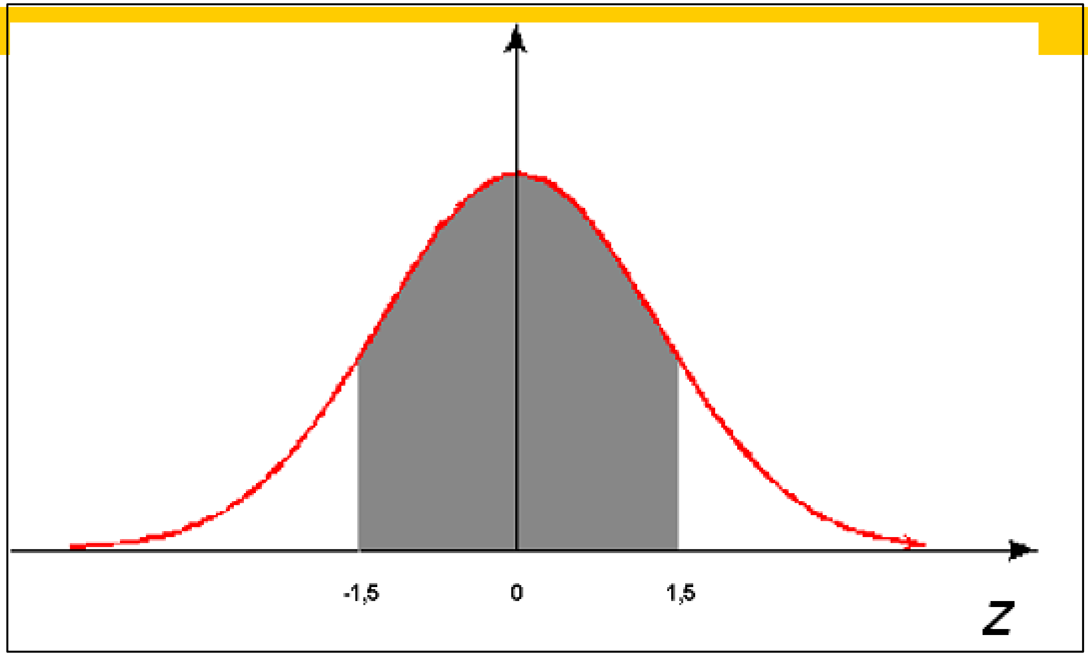


$$P(Z \leq -1,3) = P(Z \geq 1,3) = 1 - P(Z \leq 1,3) = 1 - A(1,3)$$

$$= 1 - 0,9032 = 0,0968.$$

Obs.: Pela simetria, $P(Z \leq -1,3) = P(Z \geq 1,3)$.

f) $P(-1,5 \le Z \le 1,5)$



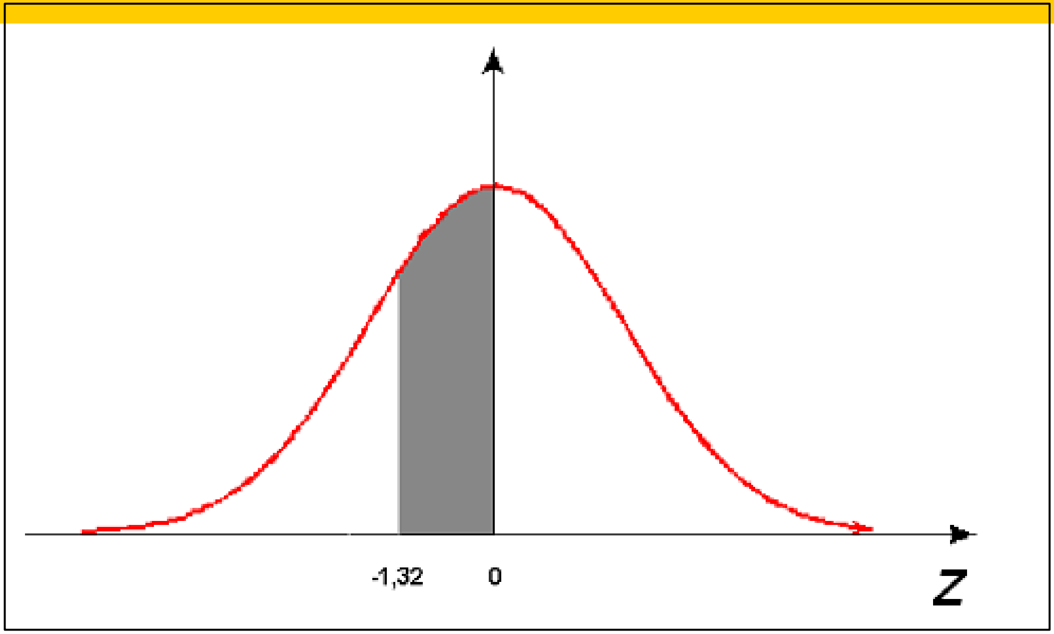
$$P(-1,5 \le Z \le 1,5) = P(Z \le 1,5) - P(Z \le -1,5)$$

$$= P(Z \le 1,5) - P(Z \ge 1,5) \;\; = P(Z \le 1,5) - [1 - P(Z \le 1,5)]$$

$$= 2 \times P(Z \le 1,5) - 1 = 2 \times A(1,5) \; - 1$$

$$= 2 \times 0,9332 \; - 1 = 0,8664.$$

g) $P(-1{,}32 < Z < 0)$



$$P(-1{,}32 < Z < 0) = P(0 < Z < 1{,}32)$$

$$= P(Z \leq 1{,}32) - P(Z \leq 0) = A(1{,}32) - 0{,}5$$

$$= 0{,}9066 - 0{,}5 = 0{,}4066.$$

Distribuição Normal-Exemplo

- QI~N(100,225)
  - Z=(QI-100)/15~N(0,1)
  - Qual a probabilidade que uma pessoa escolhida aleatoriamente tenha o QI superior a 135?

    Z=(135-100)/15=2,33

    P(Z>2.33) = 0,01 (tabela normal padrão)

  - Qual a probabilidade que uma pessoa escolhida aleatoriamente tenha o QI inferior a 90?

    Z=(90-100)/15=-0,67

    P(Z<-0,67)=P(Z>0,67)=0,2514

    - Lembre-se da simetria

  - Probabilidades que uma pessoa escolhida aleatoriamente tenha o QI entre dois valores também podem ser determinadas.

Faixa de Normalidade

- média aritmética ± desvio-padrão

- corresponde à aproximadamente 68% dos indivíduos da amostra

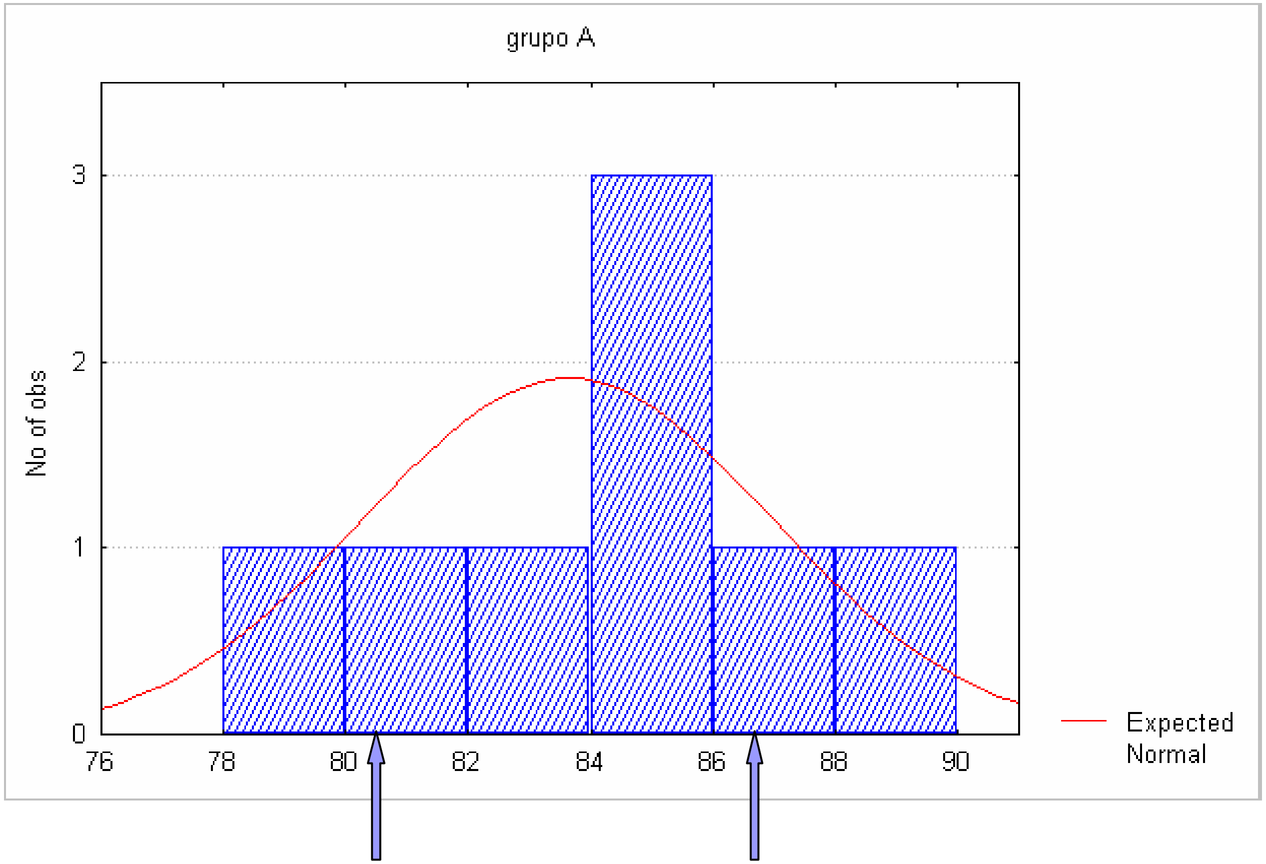
Exemplo

- **Os dados abaixo referem-se aos pesos dos pacientes em dois grupos:**

	Grupo A	*Grupo B*
	78	65
	80	69
	82	78
	85	85
	85	85
	85	93
	86	96
	88	98
Soma	669	669
Média	83,6	83,6
Mediana	85	85
Moda	85	85
N	8	8

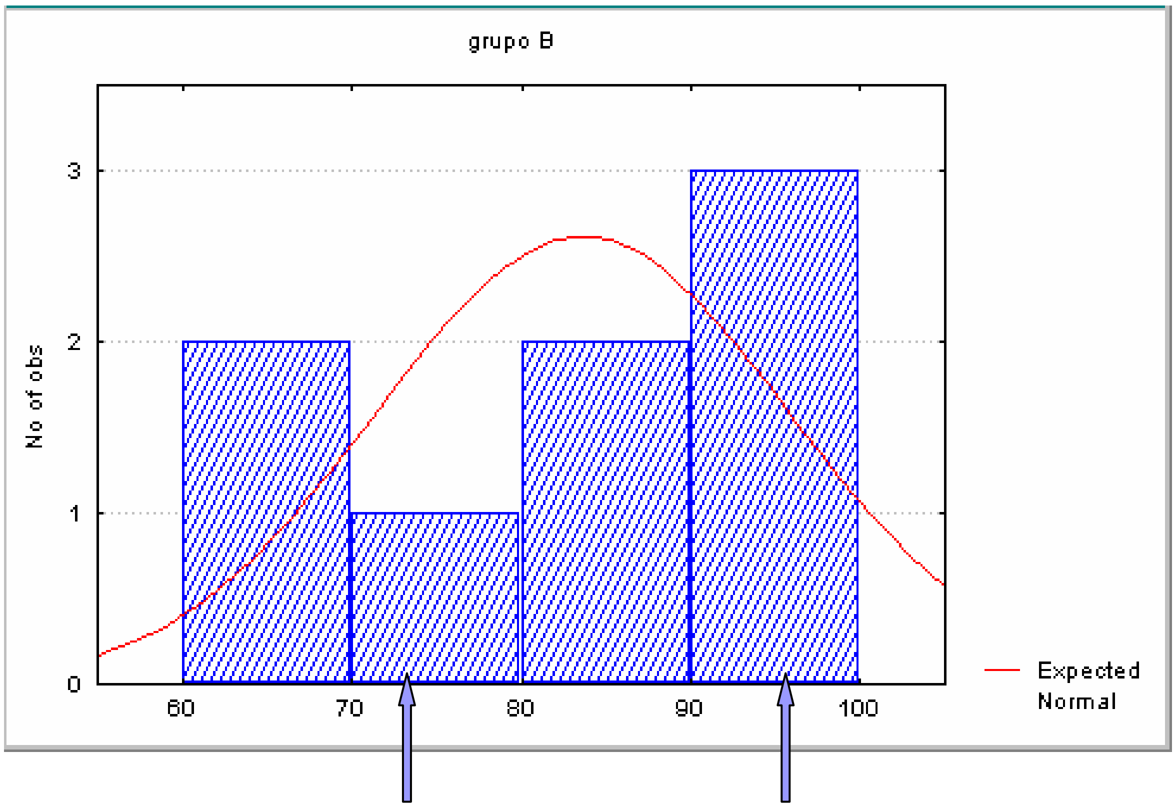
Faixa Normalidade: GRUPO A

- Limite inferior = 83,6 – 3,3 = 80,3
- Limite superior = 83,6 + 3,3 = 86,9



Faixa Normalidade: GRUPO B

- Limite inferior = 83,6 – 12,2 = 71,4
- Limite superior = 83,6 + 12,2 = 95,8



Análise Bivariada

- Muitas vezes queremos verificar se há uma relação entre duas variáveis (se as variáveis são dependentes ou não).
- Podemos construir tabelas de freqüência com dupla entrada. Essas tabelas de dados cruzados são conhecidas por tabelas de contingência, e são utilizadas para estudar a relação entre duas variáveis categóricas.

Tabela de Contingência

TABELA 4. Tipo de parto segundo categoria de internação em nascidos vivos de parto único. São Luís - MA, 1997/98

Tipo de parto	Categoria de internação					
	Pública		Privada		Total	
	f	%	f	%	f	%
Cesáreo	572	26,31	252	93,68	824	33,73
Vaginal	1602	73,69	17	6,32	1619	66,27
Total	2174	100,00	269	100,00	2443	100,00

Fonte: Silva et al (2001)

Gráficos: Duas Variáveis Qualitativas

Gráfico de barras

FIGURA 5: Óbitos por acidentes, segundo tipo e sexo.

Município de São Paulo, 1980.

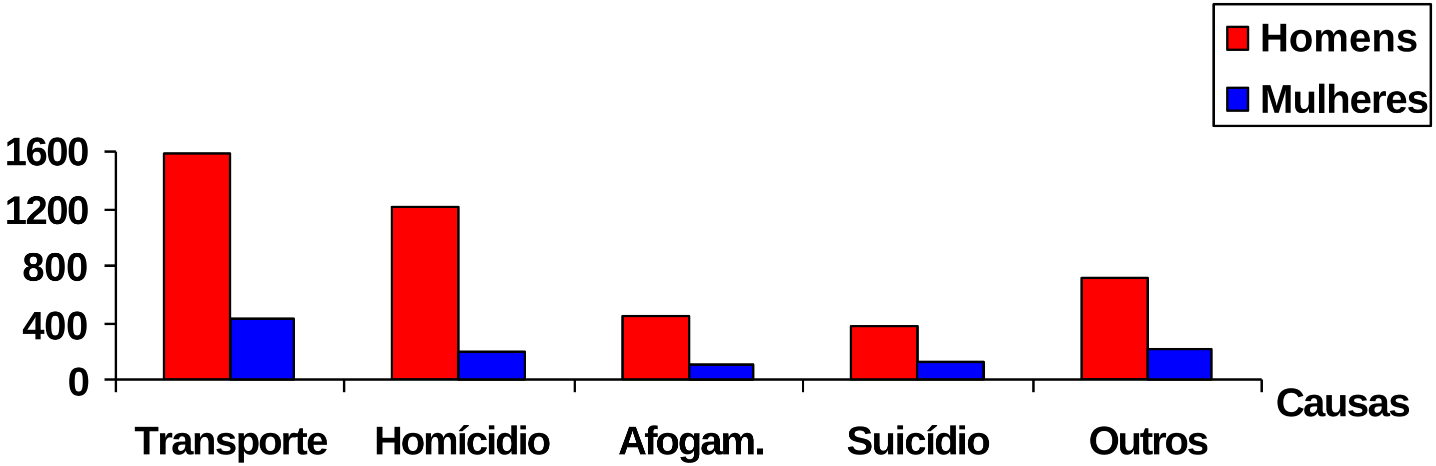


Gráfico: Duas Variáveis Quantitativas

Gráfico de Dispersão