Distribuição de Freqüência

Representação do conjunto de dados

- Distribuições de freqüência
  - Freqüência relativa
  - Freqüência acumulada

- Representação Gráfica
  - Histogramas

Organização dos dados

- Os métodos utilizados para *organizar* dados compreendem o arranjo desses dados em subconjuntos que apresentem características similares.
  - mesma idade (ou "*faixa etária*"), mesma finalidade, mesma escola, mesmo bairro, etc
- Os *dados agrupados* podem ser resumidos em tabelas ou gráficos e, a partir desses, podemos obter as estatísticas descritivas já definidas: média, mediana, desvio, etc.
- Dados organizados em grupos ou categorias/classes são usualmente designados *"distribuição de freqüência".*

Distribuição de frequência

- Uma *distribuição de freqüência* é um método de se agrupar dados em classes de modo a fornecer a quantidade (e/ou a percentagem) de dados em cada classe

- Com isso, podemos *resumir e visualizar* um conjunto de dados sem precisar levar em conta os valores individuais.

- Uma *distribuição de freqüência* (*absoluta* ou *relativa*) pode ser apresentada em tabelas ou gráficos

Distribuição de frequência

Uma distribuição de freqüência agrupa os dados por classes de ocorrência, resumindo a análise de conjunto de dados grandes.

Construindo uma distribuição de freqüência

- Adotemos o conjunto de dados que represente a população
- Ordene em ordem crescente ou decrescente

Eventos	Altura
Aluno 1	1,60
Aluno 2	1,69
Aluno 3	1,72
Aluno 4	1,73
Aluno 5	1,73
Aluno 6	1,74
Aluno 7	1,75
Aluno 8	1,75
Aluno 9	1,75
Aluno 10	1,75
Aluno 11	1,75
Aluno 12	1,76
Aluno 13	1,78
Aluno 14	1,80
Aluno 15	1,82
Aluno 16	1,82
Aluno 17	1,84
Aluno 18	1,88

Construindo uma distribuição de freqüência

- Determine a Quantidade de classes (k)
  - Regra de Sturges (Regra do Logaritmo)
    - k = 1 + 3,3log(n)
  - Regra da Potência de 2
    - k = menor valor interiro tal que $2^k \geq n$
  - Regra da Raiz Quadrada
    - $k = \sqrt{n}$
  - Bom senso !!!
    - Decida a quantidade de classes que GARANTA observar como os valores se distribuem.

Construindo uma distribuição de freqüência

Regra de Sturges (Logaritmo)	
Quantidade de dados (n)	Quantidade de Classes (k)
1	1
2	2
3 a 5	3
6 a 11	4
12 a 23	5
24 a 46	6
47 a 93	7
94 a 187	8
188 a 376	9
377 a 756	10

Regra da Potência de 2	
Quantidade de dados (n)	Quantidade de Classes (k)
1 e 2	1
3 e 4	2
5 a 8	3
9 a 16	4
17 a 32	5
33 a 64	6
65 a 128	7
129 a 256	8
257 a 512	9
513 a 1024	10

Bom Senso		
Quantidade de dados (n)	Quantidade MÍNIMA de Classes (k)	Quantidade MÁXIMA de Classes (k)
até 50	5	10
51 a 100	8	16
101 a 200	10	20
201 a 300	12	24
301 a 500	15	30
mais de 500	20	40

Construindo uma distribuição de freqüência

- Calcule a amplitude das classes (h)
  - Calcule a amplitude do conjunto de dados
    - $L = x_{máx} – x_{mín}$
  - Calcule a amplitude (largura) da classe
    - $h = L / k$
    - Arredonde convenientemente
- Calcule os Limites das Classes
  - 1ª classe: $x_{mín}$ até $x_{mín} + h$
  - 2ª classe: $x_{mín} + h$ até $x_{mín} + 2 . h$
  - ....
  - kª classe: $x_{mín} + (k-1) . h$ até $x_{mín} + k . h$

Construindo uma distribuição de freqüência

- Limite das classes
  - Utilize a notação:
    - [x,y) – intervalo de entre x (fechado) até y (aberto)
  - Freqüentemente temos que “arredondar” a amplitude das classes e, conseqüentemente, arredondar também os limites das classes.
  - Como sugestão, podemos tentar, se possível, um ajuste simétrico nos limites das classes das pontas (i.e., primeira e última) nas quais, *usualmente*, a quantidade de dados é menor.
- Ponto médio das classes
  - $x_k = L_{inferior} + (L_{superior} – L_{inferior}) / 2$

Construindo uma distribuição de freqüência

- <u>Determinação da freqüência das classes</u>
  - Consiste em agrupar os dados em cada classe e contar os totais
- <u>Traçar o gráfico</u>
  - Dividir o eixo horizontal em tantas partes quanto for o número de classes. *<u>Sugestão</u>: deixe espaço entre o eixo vertical e a primeira classe.*
  - Identifique a maior freqüência da classe na tabela e marque esse número (ou outro um pouco maior) na extremidade do eixo vertical; divida esse eixo em algumas partes e marque os valores correspondentes
  - Desenhe um retângulo, para cada classe, com largura igual à largura da classe e com altura igual à freqüência da classe

Exemplo

- Do nosso exemplo:
  - Ordenamos os dados
  - Por Sturges, temos:
    - n=18 ; k=5 (número de classes)
  - Amplitude de classes
    - Amplitude do conjunto de dados: 1,88-1,60=0,28m
    - Amplitude de classes: 0,28/5=0,056
    - Arredondado h = 0,06m

Altura
1,60
1,69
1,72
1,73
1,73
1,74
1,75
1,75
1,75
1,75
1,75
1,76
1,78
1,80
1,82
1,82
1,84
1,88

Construindo uma tabela de freqüência

- Calcule os Limites de Classe
- Arredonde os Limites de Classe nos extremos
  - 1,9-1,88=0,02
  - Distribua o excesso:
    - 1,60-0,01; 1,88+0,01
  - Ajuste todas as classes

Amplitude	0,06
Limites inferiores	Limite superior
1,60	1,66
1,66	1,72
1,72	1,78
1,78	1,84
1,84	1,90

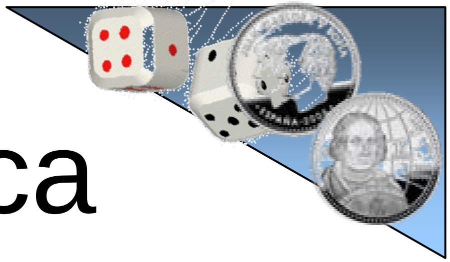
Aqui “sobra” 0,02m!

Altura
1,60
1,69
1,72
1,73
1,73
1,74
1,75
1,75
1,75
1,75
1,75
1,76
1,78
1,80
1,82
1,82
1,84
1,88

Construindo uma tabela de freqüência

- Freqüências absolutas
  - Distribua os eventos ou ocorrência por suas respectivas classes
- Freqüências acumuladas
  - Some as ocorrências de dados cumulativamente às classes
- Observação importante:
  - É muito útil representar as frequências em termos percentuais ao total de amostras

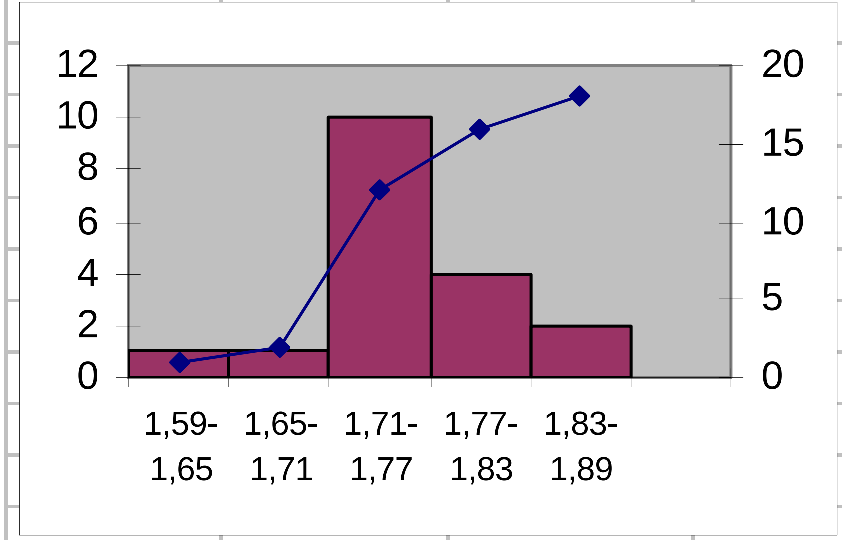
Amplitude		0,06	
Dados	Classe	Frequência	Frequência Acumulada
1,60	1,59-1,65	1	1
1,69	1,65-1,71	1	2
1,72	1,71-1,77	10	12
1,73	1,77-1,83	4	16
1,73	1,83-1,89	2	18
1,74	Total	18	
1,75			
1,75			
1,75			
1,75			
1,75			
1,76			
1,78			
1,80			
1,82			
1,82			
1,84			
1,88			



Representação Gráfica

- Histograma
  - Na abscissas, distribua as classes
  - Na ordenada da esquerda, as freqüências absolutas
  - Construa um gráfico de barras para as freqüências
  - Construa um gráfico de linha para a freqüência acumulada (utilize a escala da direita)

Amplitude	0,06	
Classe	Frequência	Frequência Acumulada
1,59-1,65	1	1
1,65-1,71	1	2
1,71-1,77	10	12
1,77-1,83	4	16
1,83-1,89	2	18
Total	18	



Distribuição de Freqüência: Histogramas e Polígonos de Freqüência

- Uma distribuição de freqüência representada por um gráfico de barras é denominada *histograma*
- Outro gráfico de interesse é o chamado *polígono de freqüência*
- O *polígono de freqüência* é obtido unindo-se os pontos médios da parte superior de cada retângulo do *histograma* com segmentos de reta
- É importante notar que tanto o histograma quanto o polígono de freqüência indicam a *freqüência absoluta* de cada classe

Distribuição de Freqüência: Histogramas e Polígonos de Freqüência

- Digamos que temos histogramas para as alturas dos estudantes de duas turmas diferentes, *traçados de acordo com as regras descritas até agora*

- *Poderíamos sobrepor os desenhos para fazer uma análise comparativa das turmas?*

- Que cuidados devemos tomar?

Distribuição de Freqüência: Histogramas e Polígonos de Freqüência

- O “*problema*” com esta regra de construção é que o histograma construído é específico para o conjunto em análise

- Para fazermos análises comparativas de conjuntos de dados diferentes, as classes devem ser as mesmas!

- Devemos, então, utilizar algum conhecimento prévio da área em estudo para definir o intervalo *aceitável* de variação dos dados e, a partir daí, definir as classes

- Essas “*classes genéricas*” servirão para o estudo de quaisquer conjunto de dados e permitirão análises comparativas

Distribuição de Freqüência: Histogramas e Polígonos de Freqüência

- Em um histograma, as classes devem SEMPRE ter a mesma largura?

- Não necessariamente!

- Existem casos em que é mais adequado agrupar os dados em classes com larguras desiguais.

- O exemplo típico é a classificação de pessoas por faixas etárias (*infantil, juvenil, adulto, sênior, etc*). Essas faixas não têm a mesma largura.

Distribuição de Freqüência: Histogramas com Classes de Larguras Desiguais

- A representação gráfica dos dados em um histograma com classes de larguras desiguais requer a transformação dos valores de freqüência absoluta em *densidade de freqüência*.

- Isso é fundamental pois devemos manter a área dos retângulos proporcionais à freqüência da classe

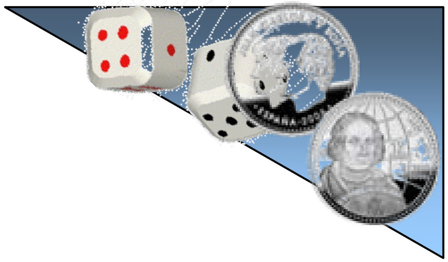
- A *densidade de freqüência* é dada por:

$$\textbf{densidade de freqüência} = \frac{\textbf{freqüência da classe}}{\textbf{largura da classe}}$$

Distribuição de Freqüência: Histogramas com Classes de Larguras Desiguais

- Isso significa que a altura das barras (*i.e., os valores na escala do eixo vertical*) NÃO representam a freqüência da classe, mas sim a densidade de freqüência.

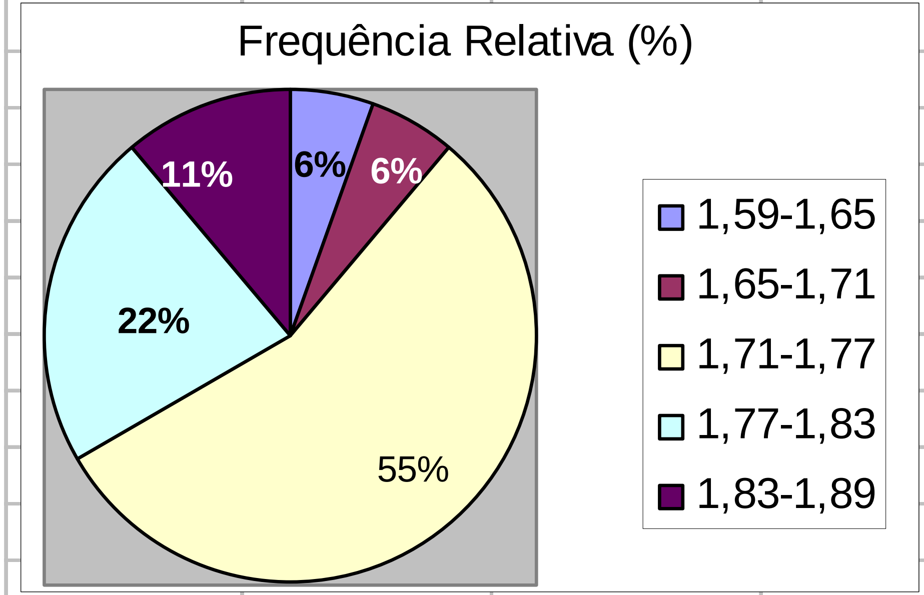
- Para calcularmos a freqüência da classe devemos multiplicar a densidade (indicada no eixo vertical) pela largura respectiva

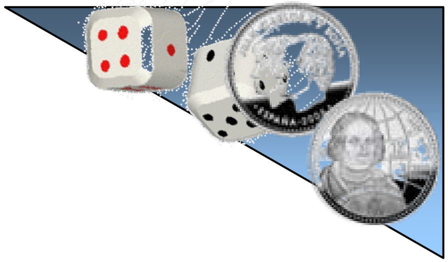


Outros Gráficos

Amplitude	0,05	
Classe	**Frequência**	**Frequência Relativa (%)**
1,59-1,65	1	6%
1,65-1,71	1	6%
1,71-1,77	10	56%
1,77-1,83	4	22%
1,83-1,89	2	11%
Total	18	

Gráfico de Pizza





Outros Gráficos

Classe	Frequência	Frequência Relativa(%)	Frequência Acumulada	Frequência Acumulada(%)
1,71-1,77	10	56%	10	56%
1,77-1,83	4	22%	14	78%
1,83-1,89	2	11%	16	89%
1,65-1,71	1	6%	17	94%
1,59-1,65	1	6%	18	100%
Total	18			

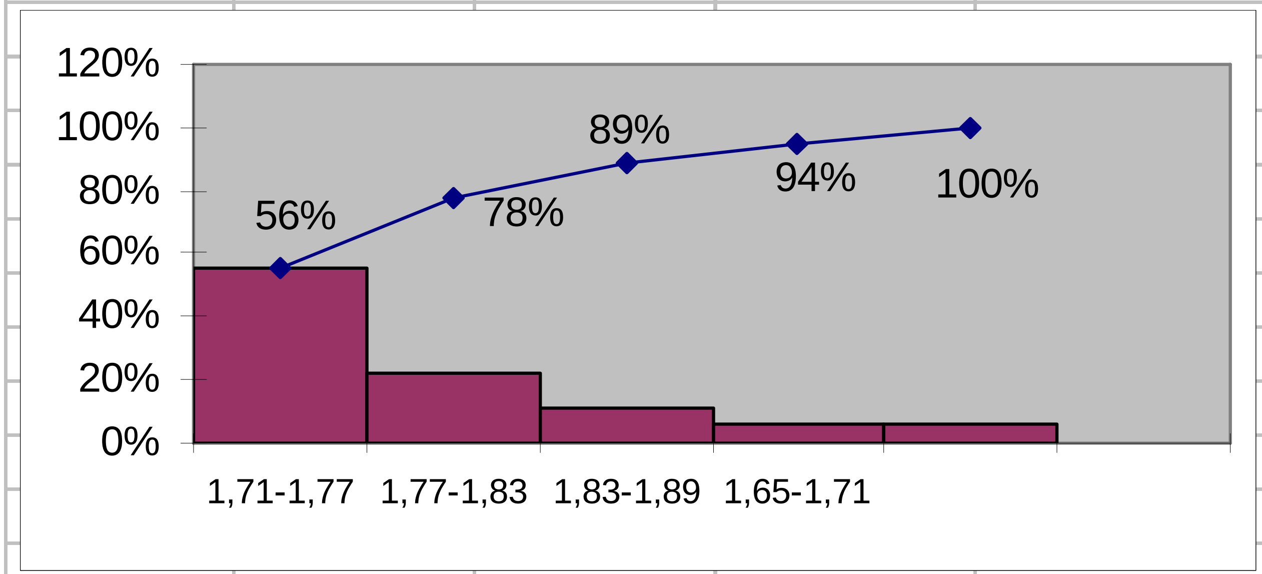
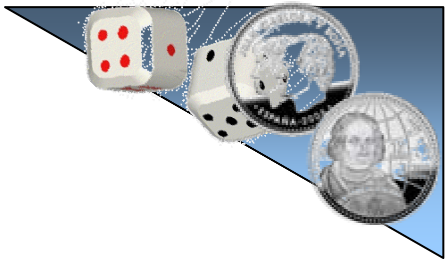
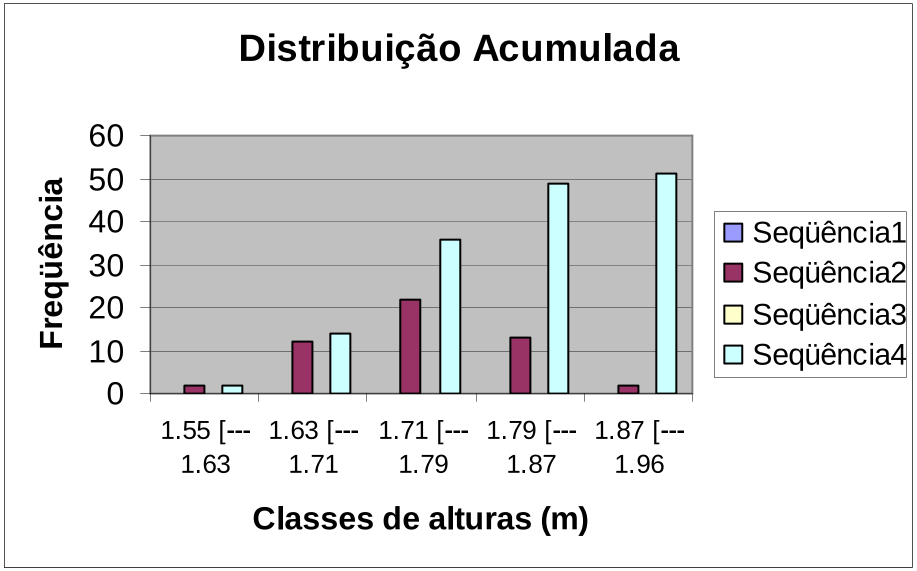


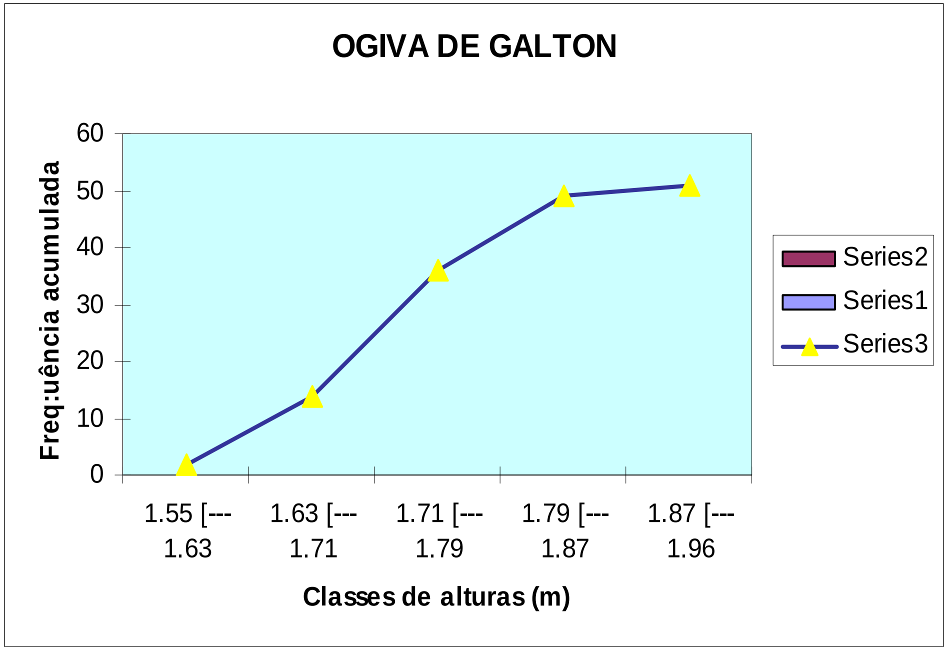
Gráfico de Pareto



Outros Gráficos

Classe de Altura (m)	Freqüência	Freq. Acumulada
1.55 [--- 1.63	2	2
1.63 [--- 1.71	12	14
1.71 [--- 1.79	22	36
1.79 [--- 1.87	13	49
1.87 [--- 1.96	2	51





Média Ponderada: Média de uma tabela de freqüência

- Quando os dados estão resumidos em uma tabela de freqüências, podemos calcular aproximadamente a média aritmética ponderando sobre:
  - Pontos médios de cada intervalo – supõe-se que todos os elementos das classes ocorrem no ponto médio das respectivas classes;
  - Exemplo: temos 7 ocorrências na faixa entre 1,75 e 1,79. Consideramos que as sete ocorrências equivalem a (1,79+1,75)/2=1,77 → ponto médio da classe.

Média Ponderada: Média de uma tabela de freqüência

$$\boxed{\bar{x} = \frac{\sum (f.x)}{\sum f}}$$

- x = ponto médio da classe
- f = *freqüência*
- Σf = n

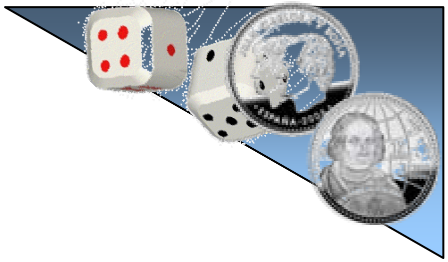
Média Ponderada

- A média ponderada é considerada “ponderada” quando os valores dos conjuntos tiverem pesos / freqüências diferentes
- Numa distribuição utilizando os valores discretos, calcula-se:

Erros por páginas	No de paginas
0	**25**
1	**20**
2	**3**
3	**1**
4	**1**

$$\bar{x} = \frac{\sum x.f}{\sum f} = \frac{\sum x.f}{n}$$

$$\bar{x} = \frac{(0 \cdot 25) + (1 \cdot 20) + (2 \cdot 3) + (3 \cdot 1) + (4 \cdot 1)}{(25 + 20 + 3 + 1 + 1)} = \frac{33}{50} = 0{,}66$$



Média Ponderada

- Quando tivermos uma distribuição com dados agrupados por classes de valores, calculamos considerando o valor de cada classe como o ponto médio respectivo da classe.

Alturas de Pessoas	Ponto Médio (Xi)	Frequência (fi)	xi.fi
1,59-1,65	1,62	**1**	**1,62**
1,65-1,71	1,68	**1**	**1,68**
1,71-1,77	1,74	**10**	**17,4**
1,77-1,83	1,80	**4**	**7,2**
1,83-1,89	1,86	**2**	**3,72**
Total		**18**	**31,62**

$$\bar{x} = \frac{\sum x.f}{\sum f} = \frac{\sum x.f}{n}$$

$$\bar{x} = \frac{31{,}62}{18} = 1{,}76$$

Cálculo da Moda para dados Agrupados

- Caso 1: dados agrupados por valores discretos → moda é o valor com maior freqüência.
- Caso 2: dados agrupados por classes
  - Moda Bruta
  - Método de King
  - Método de Czuber
  - Método de Pearson

Cálculo da Moda para dados Agrupados: Moda Bruta

- Moda Bruta
  - Tome a classe que apresenta a maior freqüência → *classe modal*
  - A moda será o ponto médio da classe modal: $(\lim_{inf} + \lim_{sup})/2$

Cálculo da Moda para dados Agrupados: King

- Método de King:

$$M_o = \lim_{\text{inf}} + \frac{f_{post}}{f_{ant} + f_{post}} h$$

- Onde
  - Lim_{inf}: limite inferior da classe modal
  - f_{ant}: freqüência da classe anterior à modal
  - f_{post}: freqüência da classe posterior à modal
  - h: amplitude da classe modal

Cálculo da Moda para dados Agrupados: Czuber

- Método de Czuber (mais preciso):

$$M_o = \lim\nolimits_{\inf} + \frac{f_{Mo} - f_{ant}}{f_{Mo} - \left(f_{ant} + f_{post}\right)} h$$

- Onde
  - Lim_{inf}: limite inferior da classe modal
  - f_{Mo}: freqüência da classe modal
  - f_{ant}: freqüência da classe anterior à modal
  - f_{post}: freqüência da classe posterior à modal
  - h: amplitude da classe modal

Cálculo da Moda para dados Agrupados: Pearson

- Método de Pearson:

$$M_o = 3M_d - 2\bar{X}$$

- Onde
  - M_d: Mediana
  - $\bar{X}$: Média

Cálculo da Mediana para dados Agrupados

- Dados agrupados por classes
  - Mediana é o valor localizado a $L_x = n/2$
  - Após cálculo de L_x, determina-se o valor da mediana por:

$$\boxed{\tilde{X} = Lim_{\inf} + \frac{h.(L_x - F_{ant})}{f_i}}$$

  - Onde:
    - L_x → Localização (posição) da Mediana
    - Fant → freqüência acumulada até a classe anterior à classe da mediana
    - f_i → freqüência absoluta da classe da mediana
    - h → amplitude de classe
    - Lim_{inf} → Limite inferior da classe da mediana

Cálculo dos Percentis para dados Agrupados por Classes

- O percentil é o valor localizado a L_{Px}=(K/100)*n
  - Onde K é o percentil desejado (ex.: $P_{45} \rightarrow$ K=45)
- Após cálculo de L_{Px}, determina-se o valor do percentil por:

$$\boxed{P_x = Lim_{\inf} + \frac{h.(L_{Px} - F_{ant})}{f_i}}$$

- Onde:
  - $L_{Px} \rightarrow$ Localização (posição) do Percentil
  - Fant $\rightarrow$ freqüência acumulada até a classe anterior à classe do percentil
  - $f_i \rightarrow$ freqüência absoluta da classe do percentil
  - h $\rightarrow$ amplitude de classe
  - $Lim_{\inf} \rightarrow$ Limite inferior da classe do percentil

Medidas de Posição Dados Agrupados: Mediana / Separatrizes (alternativo)

- Para definirmos um procedimento alternativo de cálculo da *mediana* e quaisquer outras *separatrizes*, utilizaremos o exemplo abaixo:

TABELA 1.6 – Distribuição de freqüência da variável S = salário dos empregados da seção de orçamento da Companhia Milsa.

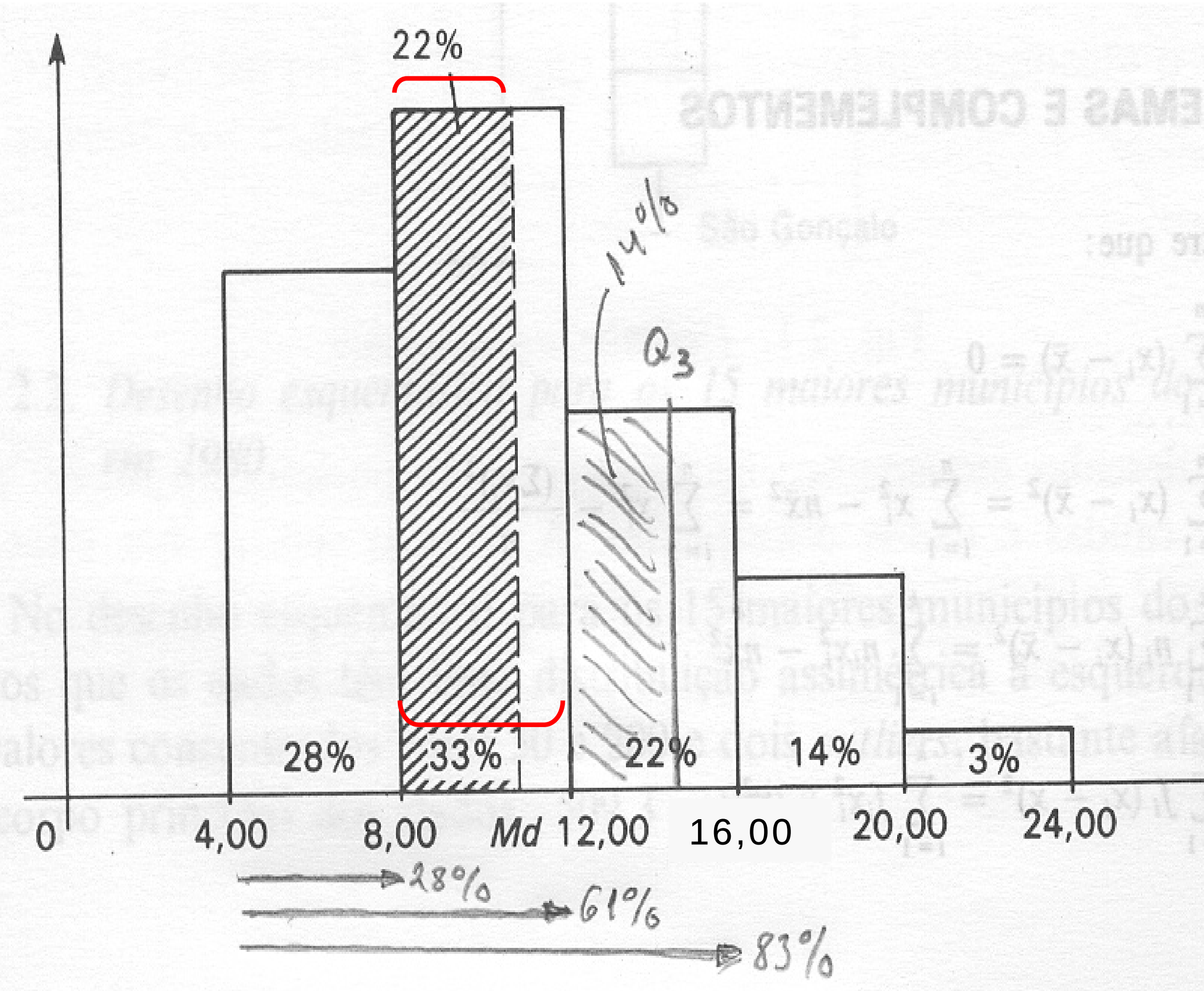
Classe de salários	*Ponto médio* s_i	*Freqüência* n_i	*Porcentagem* $100 \cdot f_i$
4,00 ⊢ 8,00	6,00	10	27,78
8,00 ⊢ 12,00	10,00	12	33,33
12,00 ⊢ 16,00	14,00	8	22,22
16,00 ⊢ 20,00	18,00	5	13,89
20,00 ⊢ 24,00	22,00	1	2,78
TOTAL	—	36	100,00

Medidas de Posição Dados Agrupados: Mediana / Separatrizes

- Encontra-se a classe onde está a *mediana.* Faz-se, então, a proporcionalidade entre a área e a base do retângulo hachurado e o que define a classe onde está a mediana

$$\frac{12{,}00 - 8{,}00}{33\%} = \frac{M_d - 8{,}00}{22\%}$$

- $M_d = 10{,}67$

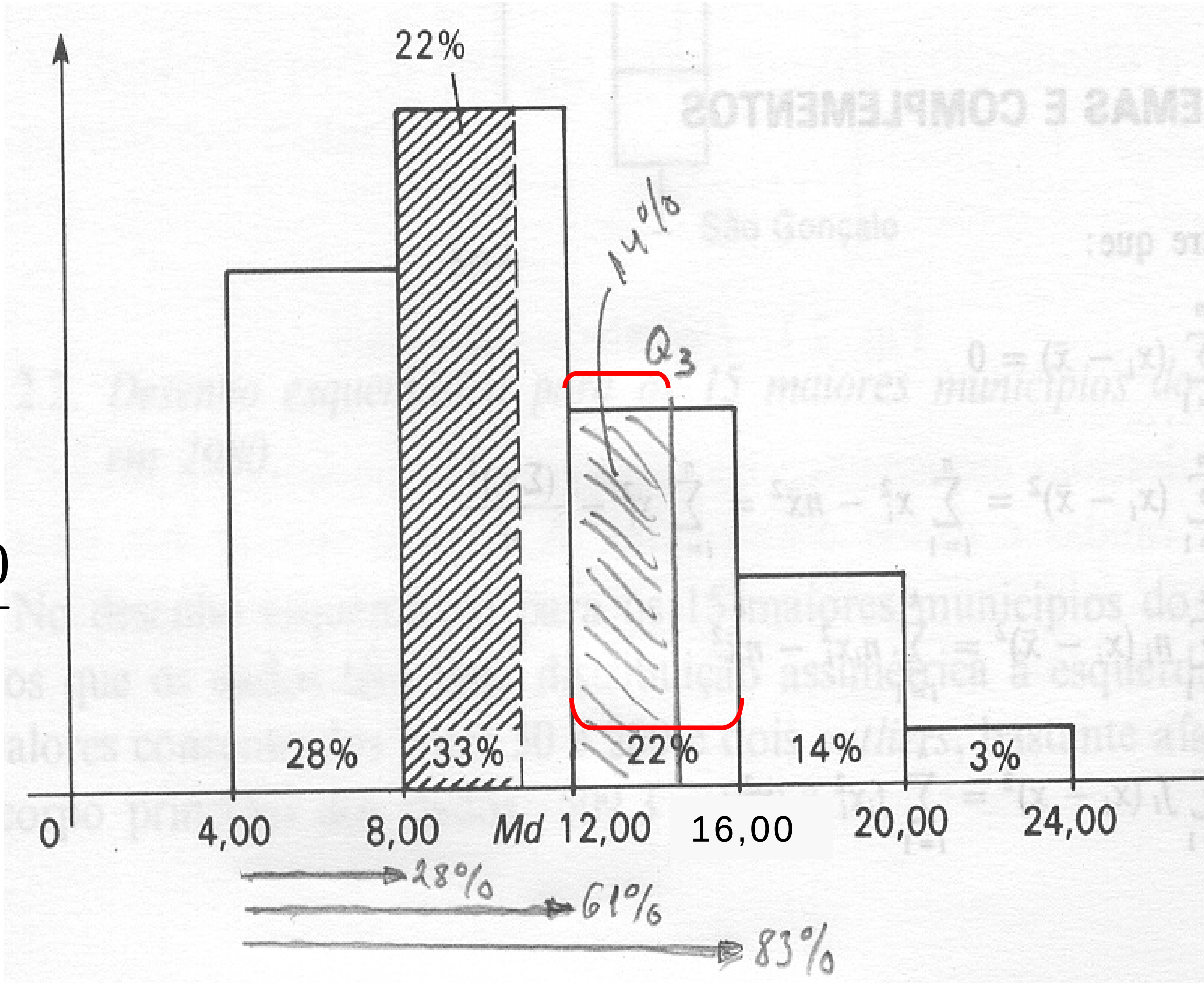


Medidas de Posição Dados Agrupados: Mediana / Separatrizes

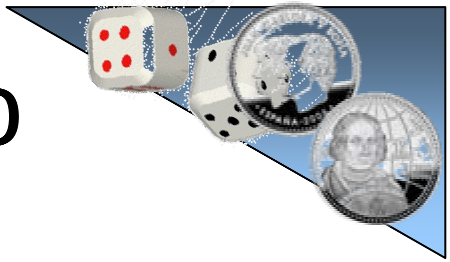
- Encontra-se a classe onde está *Q3.* Faz-se, então, a proporcionalidade entre a área e a base do retângulo hachurado e o que define a classe de *Q3*

$$\frac{18{,}00-12{,}00}{22\%}=\frac{Q_3-12{,}00}{14\%}$$

- $Q_3 = 15{,}82$



Medidas de Dispersão (Dados Agrupados)

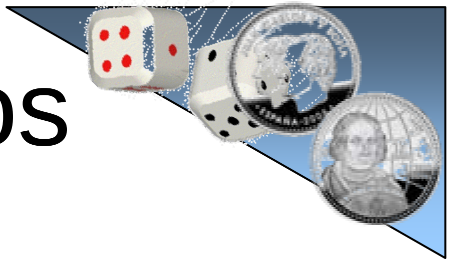


- O *desvio-padrão,* nesse caso, faz uma *ponderação* da distância dos pontos médios de cada classe para a média, e a respectiva freqüência de valores:

$$s = \sqrt{\frac{\sum_{j=1}^{k} f_j \left(\tilde{x}_j - \bar{x}\right)^2}{n-1}}$$ ***(amostra)***

$$\sigma = \sqrt{\frac{\sum_{j=1}^{k} f_j \left(\tilde{x}_j - \mu\right)^2}{N}}$$ ***(população)***

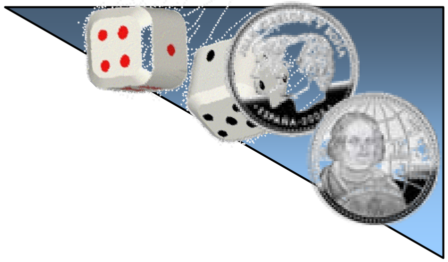
Desvio padrão de dados agrupados



$$s = \sqrt{\frac{n[\sum(f.x^2)] - [\sum(f.x)]^2}{n(n-1)}}$$

Desvio padrão para uma tabela de freqüências

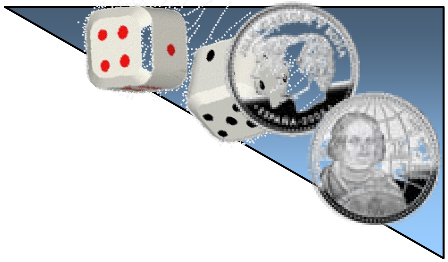
- x = ponto médio da classe
- f = frequência da classe
- n = tamanho da amostra (ou Σ*f* = soma das frequências)

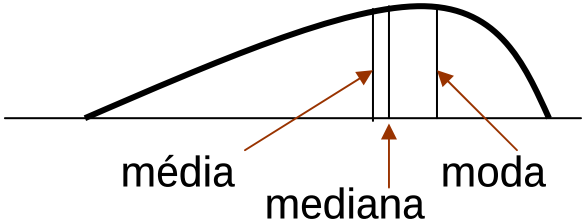


Assimetria

- Comparando a média, a moda e a mediana, podemos concluir pela assimetria da distribuição:
  - Assimetria: não simetria – distribuição tende mais para um lado
- Dados negativamente assimétricos (assimetria para a esquerda)
  - Média e mediana à esquerda da moda
  - Em geral, média à esquerda da mediana
- Dados positivamente assimétricos (assimetria para a direita)
  - Média e mediana à direita da moda
  - Em geral, média à direita da mediana

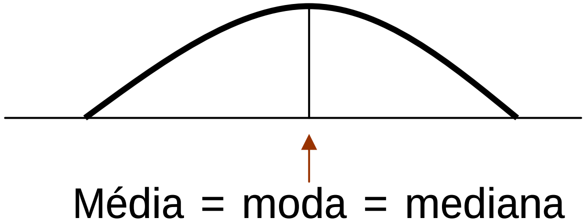
Assimetria





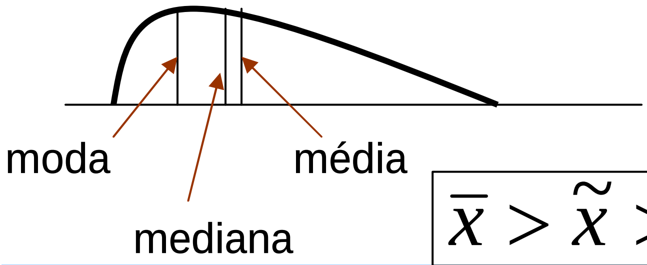
Assimétrica à esquerda

$$\bar{x} < \tilde{x} < Mo$$



Simétrica

$$\bar{x} = \tilde{x} = Mo$$



$$\bar{x} > \tilde{x} > Mo$$

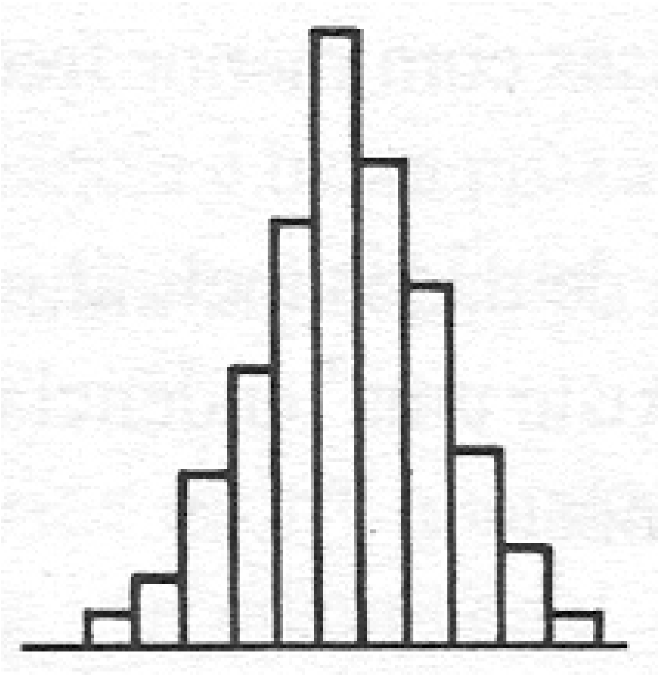
Assimétrica à direita

Interpretando Histogramas

- Histograma é uma ferramenta estatística que permite resumir informações de um conjunto de dados, visualizando a *forma da distribuição* desses dados, a *localização* do valor central e a *dispersão* dos dados em torno do valor central

- Ou seja, em análises de processos produtivos, freqüentemente obtemos informações úteis sobre a população/amostra de dados coletados pela análise da *forma do histograma*

Simétrico ou em Forma de Sino

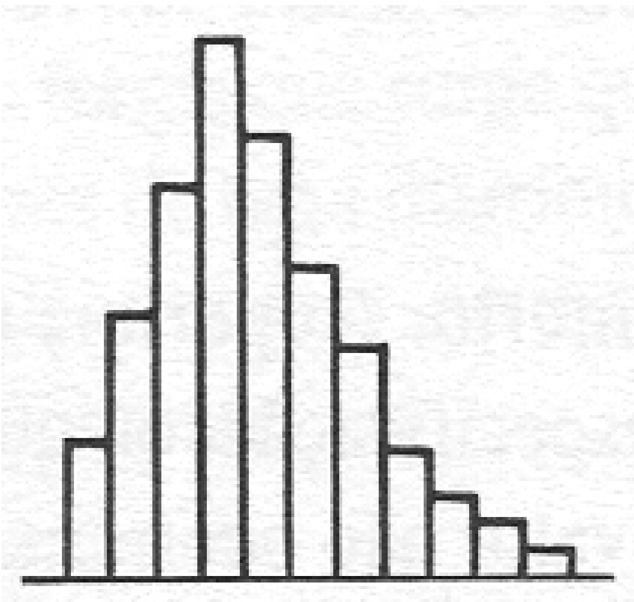


- O valor médio está localizado no centro do histograma
- A freqüência é mais alta no meio e diminui gradualmente na direção dos extremos

- Ocorre quando não existem restrições aos valores que a *variável de controle* pode assumir

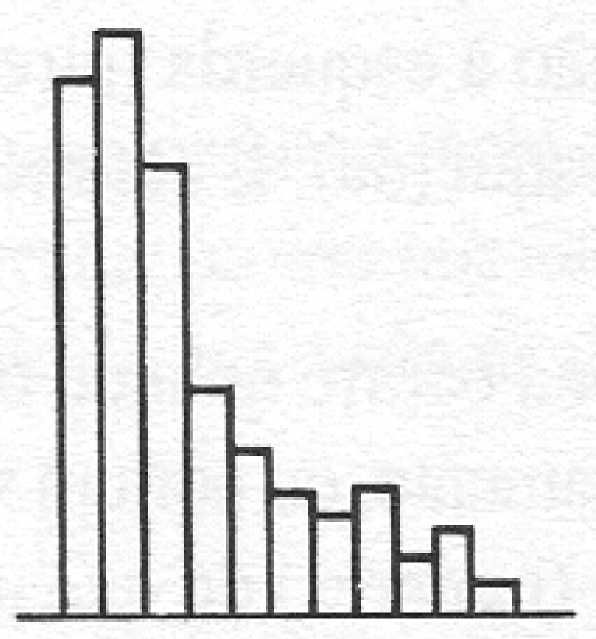
- *Processo* geralmente sob controle, somente causas comuns estão presentes
- *Processo* usualmente está estável

Assimétrico



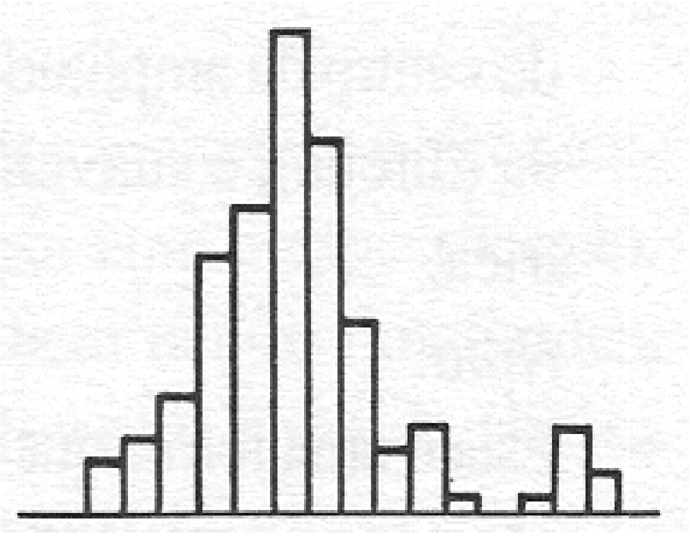
- O valor médio está localizado fora do centro do histograma
- A freqüência diminui gradativamente em um dos lados e de modo um tanto abrupto do outro lado
- Ocorre quando não é possível que a *variável de controle* assuma valores mais altos (ou mais baixos)
- *Processo* em que o limite inferior (superior) é controlado *(apenas um limite de especificação)*
- Por exemplo, teoricamente é impossível valores inferiores à 0% para a *variável impureza*

Despinhadeiro



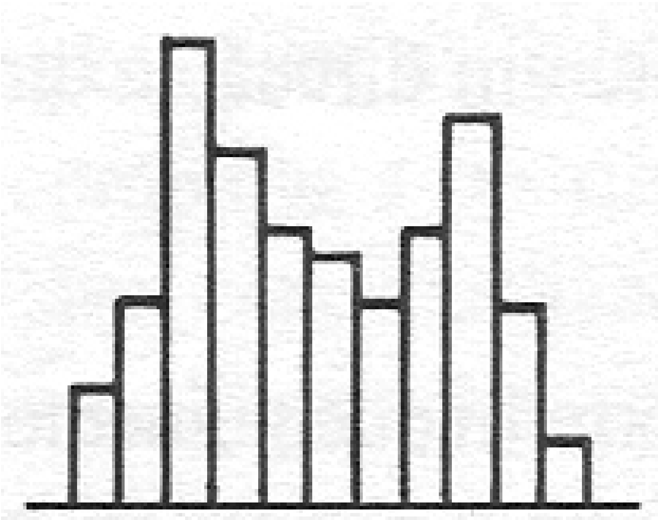
- O valor médio está localizado fora do centro do histograma
- A freqüência diminui abruptamente de um dos lados e suavemente em direção ao outro
- *Processo* não atende às especificações e uma inspeção 100% é realizada para eliminar produtos defeituosos

Ilhas Isoladas ou Pico Isolado



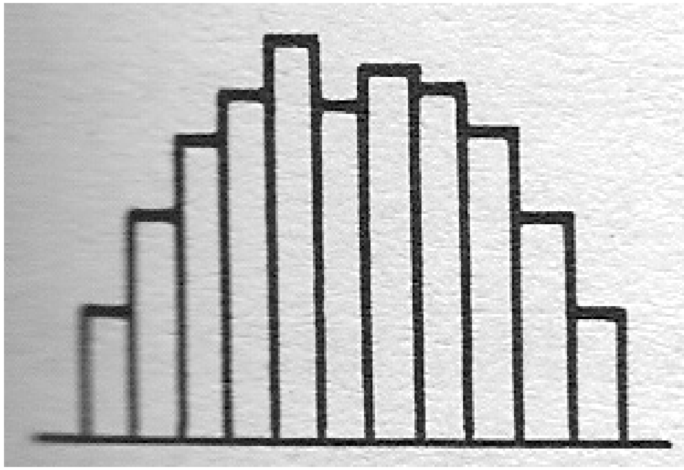
- Parte do gráfico é *relativamente* simétrica com o acréscimo de algumas classes mais afastadas de menores freqüências
- Ocorre quando dados de outra distribuição, diferente da distribuição da maior parte das medidas, são incluídos
- *Processo* com anormalidades, ou erro de medição e/ou registro de dados, ou inclusão de dados de um processo diferente

Bimodal ou com Dois Picos



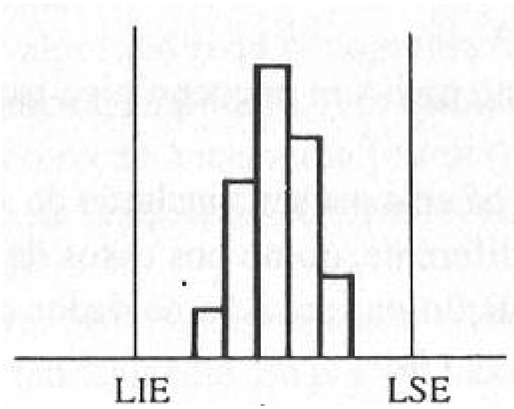
- A freqüência é mais baixa no centro do histograma e existe um “pico” em cada lado
- Ocorre quando dados de duas distribuições, com médias muito diferentes, são misturados
- Os valores da *variável de controle* devem estar associados a duas máquinas ou dois turnos distintos, por exemplo

Achatado ou Platô

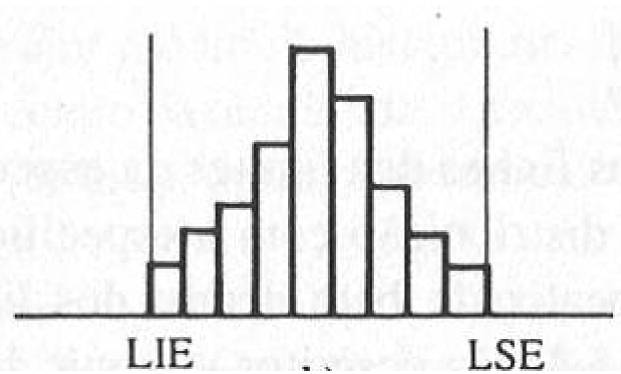


- Todas as classes possuem mais ou menos a mesma freqüência, exceto aquelas das extremidades
- Ocorre quando dados de duas distribuições, com médias *<u>não muito</u>* diferentes, são misturados
- Os valores da *variável de controle* devem estar associados a níveis distintos de algum (ou alguns) dos fatores que constituem o processo em análise

Histogramas e Limites de Especificação de Processos

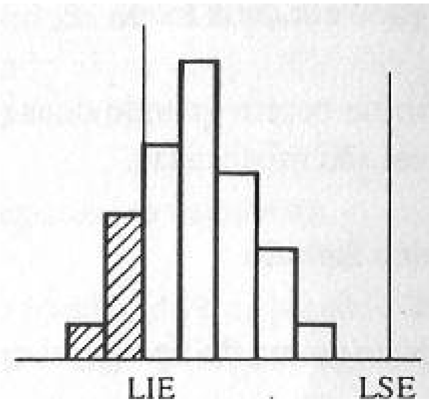


- <u>Atende, com folga, os limites de especificação</u>
- Média no centro da faixa de especificação
- Variabilidade aceitável
- Manter a situação atual

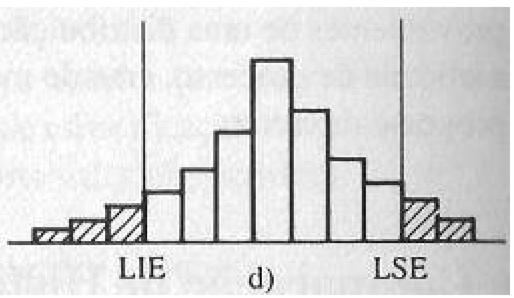


- <u>Especificação atendida sem nenhuma margem extra</u>
- Média no centro da faixa de especificação
- Variabilidade um pouco elevada
- Adotar medidas para reduzir um pouco a variabilidade

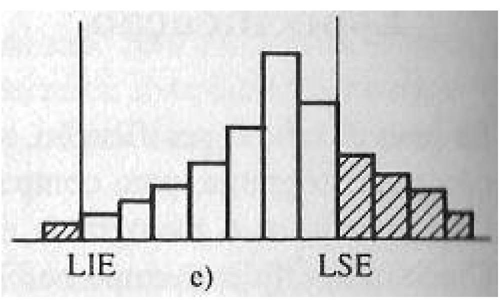
Histogramas e Limites de Especificação de Processos



- *Não atende* os limites de especificação
- Média deslocada para a esquerda
- Variabilidade aceitável
- Adotar medidas para deslocar a média para o centro (valor nominal)



- *Não atende* os limites de especificação
- Média no centro da faixa de especificação
- Variabilidade elevada
- Adotar medidas para reduzir a variabilidade



- *Não atende* os limites de especificação
- Média deslocada para a esquerda
- Variabilidade elevada
- Adotar medidas para deslocar a média para o centro e reduzir a variabilidade

Coeficiente de Assimetria

Coeficiente de Assimetria de Pearson (As)

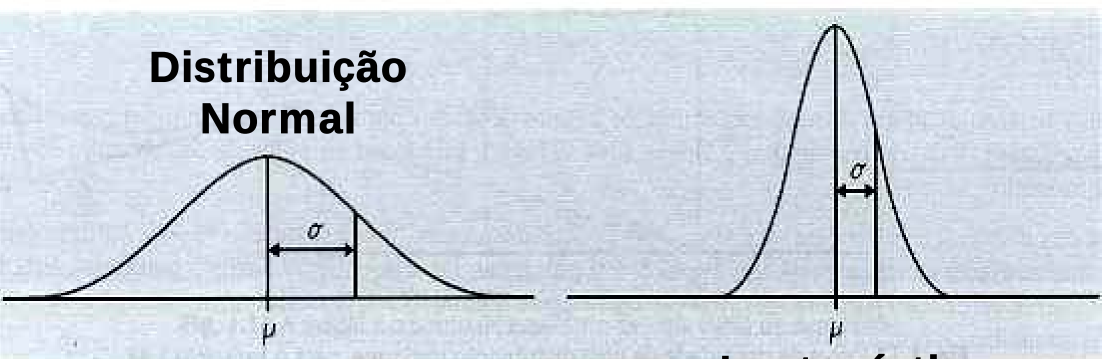
$$\boxed{As = \frac{3.(\bar{x} - \tilde{x})}{s}}$$

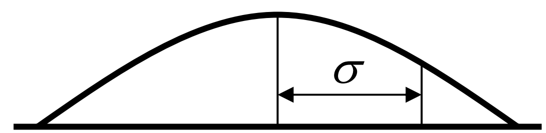
- Permite comparar duas ou mais distribuições diferentes e avaliar qual é mais assimétrica.
- Quanto maior o Coeficiente de Assimetria de Pearson, mais assimétrica é curva.
  - Assimétrica moderada: $0,15 < |As| < 1$
  - Assimétrica forte: $|As| > 1$

Curtose

Grau de achatamento (ou afilamento) de uma distribuição em relação com a distribuição normal.

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





Mesocúrtica

C=0,263

Leptocúrtica

C<0,263

Platicúrtica

C>0,263