PROF.FERNANDO MACEDO - **Profº. CEFETCE**

ESTATÍSTICA APLICADA

PARTE I - ESTATÍSTICA DESCRITIVA

Introdução

A Significância e a Abrangência da Estatística
Porque a estatística é importante?

Os métodos estatísticos são usados hoje em quase todos os campos de investigação científica, já que eles capacitam-nos a responder a um vasto número de questões, tais como as listadas abaixo:

1) Como os cientistas avaliam a validade de novas teorias?
2) Como os pesquisadores médicos testam a eficiência de novas drogas ?
3) Como os demógrafos prevêem o tamanho da população do mundo em qualquer tempo futuro?
4) Como pode um economista verificar se a mudança atual no Índice de Preços ao Consumidor é a continuação de uma tendência secular, ou simplesmente um desvio aleatório?
5) Como é possível para alguém predizer o resultado de uma eleição entrevistando apenas algumas centenas de eleitores ?

O que é Estatística ?

A noção de "Estatística" foi originalmente derivada da mesma raiz da palavra "Estado", já que foi a função tradicional de governos centrais no sentido de armazenar registros da população, nascimentos e mortes, produção das lavouras, taxas e muitas outras espécies de informação e atividades. A contagem e mensuração dessas quantidades gera todos os tipos de dados numéricos que são úteis para o desenvolvimento de muitos tipos de funções governamentais e formulação de políticas públicas.

A Estatística, como um método científico, refere-se ao projeto de experimentos e a descrição e interpretação de observações que são feitas. De um ponto de vista moderno, a Estatística é freqüentemente definida como um método de tomada de decisão em face da aleatoriedade dos fenômenos. Em uma mais vasta perspectiva, o escopo da estatística pode ser pensado em termos de três áreas diferentes de estudos: (1) a Estatística Descritiva (2) A Estatística Indutiva e (3) A Teoria da Decisão Estatística.

Estatística Descritiva

A estatística Descritiva refere-se ao corpo de métodos desenvolvidos para coletar, organizar, apresentar e descrever dados numéricos. Essa área da Estatística refere-se às seguintes tarefas:

1) Encontrar um método apropriado de coletar dados numéricos eficientemente e acuradamente para um dado problema.

2) Determinar um formato eficiente , tal como uma apresentação tabular, para a organização dos dados de uma forma sistemática e ordenada, de maneira que a informação fornecida pelos dados possa ser observada com grande facilidade e precisão.
3) Apresentar dados numéricos, seja organizados ou não, de forma que as características e o comportamento dos dados são clara e facilmente revelados. Tais apresentações São feitas por meio de métodos gráficos.
4) Sumarizar ou descrever cada característica ou propriedade dos dados por um simples número, tal como uma média, uma porcentagem ou alguma outra medida apropriada, a qual é calculada a partir dos dados por meio de uma fórmula derivada a partir de algum princípio válido.

Estatística Indutiva

A Estatística Indutiva, que é também freqüentemente chamada de inferência estatística ou estatística inferencial, em contraste com a estatística descritiva, é essencialmente analítica em sua natureza. Consiste de um conjunto de princípios ou teoremas que nos permitem generalizar acerca de alguma característica de uma "população" a partir das características observadas de uma "amostra". Nessa definição, uma *população* é o conjunto de todos os itens, objetos, coisas ou pessoas a respeito das quais a informação é desejada para a solução de um problema. Uma *amostra* é um grupo de itens selecionados por um método cuidadosamente concebido e projetado a partir de uma população. Existem diferentes tipos de amostras, dependendo dos diferentes métodos de seleção disponíveis. Uma amostra aleatória simples, falando em termos simplificados, é aquela que é selecionada de tal forma que cada e todos os itens na população tem a mesma chance de serem incluídos na amostra.

Se uma medida descritiva é calculada a partir dos dados da população ela é chamada de *parâmetro populacional*, ou simplesmente *parâmetro;* se é calculada a partir dos dados da amostra ela é chamada de *estatística amostral*, ou simplesmente *estatística.* Considerando esses conceitos podemos definir *estatística indutiva* como o processo de generalizar acerca de do valor de um parâmetro a partir do valor de uma estatística. Existem dois procedimentos de inferência distintos mas relacionados: estimação e teste de hipóteses. *Estimação* é processo de usar o valor de uma estatística amostral para estimar o valor de um parâmetro que é desconhecido, mas é uma constante. Como um exemplo suponhamos que temos uma população de 100.000 bolas de gude em um saco, todas as quais são idênticas exceto pela cor, e que não podemos vê-las embora saibamos que uma parte delas são brancas e o restante são pretas. Suponha que desejamos ter uma idéia da proporção de, digamos, bolas brancas nessa população. Suponha que para conseguir isso selecionamos 1.000 bolas aleatoriamente do saco e verificamos que 350 são brancas. Isso significa que nossa proporção amostral de bolas brancas é 35 %. A partir disso concluímos que a proporção populacional de bolas brancas é também 35 %. Fazendo isso nós realizamos o que é chamado de *estatística pontual.*

Mas afirmar que a proporção de bolas brancas em toda a população é exatamente igual a proporção daquela amostra particular é como dar um tiro no escuro: o valor da proporção amostral é um resultado aleatório e depende de cada amostra de 1.000 bolas escolhida da população. Pode ser que por uma enorme casualidade o resultado daquela amostra que escolhemos coincida exatamente com o valor da proporção de bolas brancas em toda a população. Mas as chances de que isso não ocorra são

muito grandes. Uma forma de contornarmos esse problema é afirmarmos que as chances são de 95 em 100 (ou de 95 %) de que o intervalo formado pela proporção amostral acrescida e diminuída de 3 pontos percentuais contenha o verdadeiro valor da proporção populacional desconhecido. Ou seja, construímos um intervalo com limites 35 + 0,03 x 35 = 36,05 e 35 - 0,03 x 35 = 33,95 e afirmamos (com base em algum princípio obtido a partir da teoria estatística) que as chances são de 95 em 100 de que o verdadeiro valor da proporção populacional esteja localizado dentro desse intervalo. Quando uma afirmativa dessa natureza é feita estamos realizando o que se chama de *estimativa por intervalo*.

1. Estatística Descritiva

2.1 Tipos de Variáveis

Existem diversos tipos de variáveis que serão utilizadas em um estudo estatístico. É importante compreender o conceito matemático de variável. Variável é uma abstração que se refere a um determinado aspecto do fenômeno que está sendo estudado. Podemos afirmar que a quantidade colhida da safra anual de soja é uma variável. Representemos essa variável pela letra X. Essa variável pode assumir diversos valores específicos, dependendo do anos de safra, por exemplo, X_{1986}, X_{1990} e X_{1992}. Esses valores que a variável assume em determinados anos não são a própria variável , mas valores assumidos ela para determinados objetos ou pessoas da amostra ou da população. Se uma amostra tiver 50 indivíduos podemos referimo-nos a X como sendo a variável nota de estatística e a X_{30} como a nota de um indivíduo particular, no caso o trigésimo.

Variáveis quantitativas - referem-se a quantidades e podem ser medidas em uma escala numérica. Exemplos: idade de pessoas, preço de produtos, peso de recém nascidos.

As variáveis quantitativas subdividem-se em dois grupos: variáveis quantitativas discretas e variáveis quantitativas contínuas. Variáveis discretas são aquelas que assumem apenas determinados valores tais como 0,1,2,3,4,5,6 dando saltos de descontinuidade entre seus valores. Normalmente referem-se a contagens. Por exemplo: número de vendas diárias em uma empresa, número de pessoas por família, quantidade de doentes por hospital.[1] As variáveis quantitativas contínuas são aquelas cujos valores assumem uma faixa contínua e não apresentam saltos de descontinuidade. Exemplos dessas variáveis são o peso de pessoas, a renda familiar, o consumo mensal de energia elétrica, o preço de um produto agrícola.[2] As variáveis quantitativas contínuas referem-se ao conjunto dos números reais ou a um de seus subconjuntos contínuos.

Variáveis Qualitativas - referem-se a dados não numéricos.[3] Exemplos dessas variáveis são o sexo das pessoas, a cor, o grau de instrução.

As variáveis qualitativas subdividem-se também em dois grupos: as variáveis qualitativas ordinais e as variáveis qualitativas nominais. As variáveis qualitativas ordinais são aquelas que definem um ordenamento ou uma hierarquia. Exemplos são o grau de instrução, a classificação de um estudante no curso de estatística, as posições das 100 empresas mais lucrativas, etc. As variáveis qualitativas

---

3

nominais por sua vez não definem qualquer ordenamento ou hierarquia. São exemplos destas a cor , o sexo, o local de nascimento, etc.[4]

Dependendo da situação uma variável qualitativa pode ser representada (codificada) através de emprego de números (por exemplo: em sexo representamos homens como sendo “0” e mulheres como sendo “1”). Mas no tratamento estatístico dessa variável codificada não podemos considerá-la como sendo quantitativa. Ela continua sendo uma variável qualitativa (pois o é em sua essência e natureza) apesar de sua codificação numérica que tem como finalidade uma maior finalidade de tabulação de resultados.

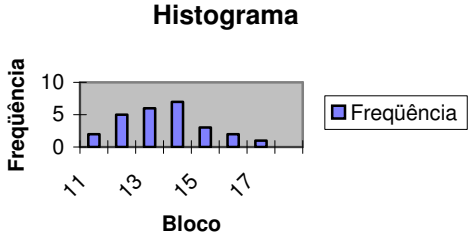
2.2 Tabelas e Distribuições de Freqüência

A análise estatística se inicia quando um conjunto de dados torna-se disponível de acordo com a definição do problema da pesquisa. Um conjunto de dados, seja de uma população ou de uma amostra contem muitas vezes um número muito grande de valores. Além disso, esses valores, na sua forma bruta, encontram-se muito desorganizados. Eles variam de um valor para outro sem qualquer ordem ou padrão. Os dados precisam então ser organizados e apresentados em uma forma sistemática e seqüencial por meio de uma tabela ou gráfico. Quando fazemos isso, as propriedades dos dados tornam-se mais aparentes e tornamo-nos capazes de determinar os métodos estatísticos mais apropriados para serem aplicados no seu estudo.

A freqüência de uma observação é o número de repetições dessa observação no conjunto de observações. A distribuição de freqüência é uma função formada por pares de valores sendo que o primeiro é o valor da observação (ou valor da variável) e o segundo é o número de repetições desse valor.

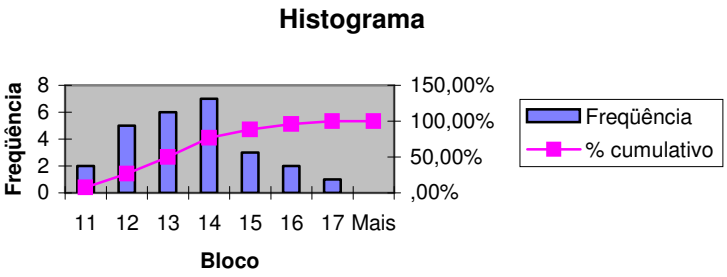
2.3 Histogramas

Histograma é uma representação gráfica de uma tabela de distribuição de freqüências. Desenhamos um par de eixos cartesianos e no eixo horizontal (abscissa) colocamos os valores da variável em estudo e no eixo vertical (ordenadas) colocamos os valores das freqüências. O histograma tanto pode ser representado para as freqüências absolutas como para as freqüências relativas. No caso do exemplo anterior, o histograma seria:



histograma de freqüência acumulada (ou ogiva) é a representação gráfica do comportamento da freqüência acumulada. Na figura abaixo a ogiva é mostrada em sobreposição ao histograma.

4



2.4 Tabulação de Freqüência e Histograma para Variáveis Contínuas

Até agora vimos como são calculadas as freqüências (relativas e acumuladas) para variáveis quantitativas discretas. Nesse caso a tabulação dos resultados é mais simples. Mas quando tratamos de variáveis quantitativas contínuas os valores observados devem ser tabulados em intervalos de classes. Para a determinação dessas classes não existe uma regra pré estabelecida, sendo necessário um pouco de tentativa e erro para a solução mais adequada. Suponhamos que as safras agrícolas de um determinado produto, em uma determinada região seja dada pela tabela a seguir:

Ano	Safra (1000 t)	Ano	Safra (1000 t)
1	280	10	365
2	305	11	280
3	320	12	375
4	330	13	380
5	310	14	400
6	340	15	371
7	310	16	390
8	340	17	400
9	369	18	370

1. **Definir o número de classes**. O número de classes não deve ser muito baixo nem muito alto. Um número de classes pequeno gera amplitudes de classes grandes o que pode causar distorções na visualização do histograma. Um número de classes grande gera amplitude de classes muito reduzidas. Foram definidas regras práticas para a determinação do número de classes, sendo que este deve variar entre 5 e 20 (5 para um número muito reduzido de observações e 20 para um número muito elevado). Se **n** representa o número de observações (na amostra ou na população, conforme for o caso) o número aproximado de classes pode ser calculado por Número de Classes = $\sqrt{n}$ arredondando os resultados. No caso do exemplo anterior temos n = 18 e $\sqrt{18} = 4{,}24$ e podemos adotar um número de 5 classes, que será razoável.

2. **Calcular a amplitude das classes**. Essa será obtida conhecendo-se o número de classes e amplitude total dos dados. A amplitude total dos dados é o resultado da subtração valor máximo - valor mínimo da série de dados. A amplitude de classe será:

$$\text{Amplitude de classe} = \frac{\text{Valor Maximo - Valor Minimo}}{\text{número de classes}}$$

Em geral, o valor do resultado é também arredondado para um número inteiro mais adequado. No nosso exemplo temos:

$$\text{Amplitude de Classe} = \frac{430 - 280}{5} = 30$$

3. **Preparar a tabela de seleção com os limites de cada classe**. Na tabela abaixo apresentamos para os dados do nosso exemplo o limite inferior e superior de cada uma das 5 classes de freqüência.

Classe	Limite inferior	Limite Superior
1	280	310
2	310	340
3	340	370
4	370	400
5	400	430

Observa-se na tabela acima que o limite superior de cada classe coincide com o limite inferior da classe seguinte. Prevendo-se que pode ocorrer que o valor de uma observação seja exatamente igual ao valor do limite de classe deve-se estabelecer um critério de inclusão. Para evitar esse tipo de dificuldade normalmente se estabelece que o limite superior de cada classe é aberto (e conseqüentemente, o limite inferior de cada classe é fechado), ou seja, cada intervalo de classe não inclui o valor de seu limite superior, com exceção da última classe.

4. **Tabular os dados por classe de freqüência**. A partir da listagem de dados seleciona-se para cada um deles qual é a sua classe de freqüência e acumula-se o total de freqüência de cada classe. De acordo com nosso exemplo, teremos:

Classe	Freqüência Absoluta Simples	Freqüência Relativa Simples
280 - 310	3	0,12 (12 %)
310 - 340	4	0,16 (16 %)
340 - 370	6	0,24 (24 %)
370 - 400	7	0,28 (28 %)
400 - 430	5	0,20 (20%)
Total	25	1,00 (100 %)

Veremos adiante, quando discutirmos as medidas de posição e de dispersão, que quando agrupamos dados numéricos em intervalos de classe ocorre perda de informação o que leva a resultados não tão precisos do que aqueles que seriam obtidos a partir dos dados originais sem agrupamento.

2.5 Medidas de Posição e de Dispersão

Podemos considerar que a Estatística Descritiva subdivide-se em duas partes. Na primeira, abordada anteriormente, são estudadas as formas de apresentação dos dados para que fiquem salientadas as suas características principais. Na segunda, que começaremos a tratar agora, abrange as medidas descritivas na forma de simples números que representam de forma sintética essas características da distribuição estatística dos dados. Estudaremos, a rigor, quatro tipos de medidas:

1. *Medidas de Tendência Central (ou medidas de posição).* Essa propriedade dos dados refere-se a localização do centro de uma distribuição. Elas nos indicam qual é a localização dos dados (no eixo que representa o conjunto dos números inteiros se estivermos tratando de uma variável quantitativa contínua).
2. *Medidas de Dispersão*. Essa propriedade revela o grau de variação dos valores individuais em torno do ponto central.
3. *Assimetria.* É a propriedade que indica a tendência de maior concentração dos dados em relação ao ponto central.
4. *Curtose*. É a característica que se refere ao grau de achatamento, ou a taxa na qual a distribuição cresce ou cai da direita para a esquerda.

2.5.1 Uma Nota sobre Notação Estatística

Utilizaremos as letras maiúsculas para representar as variáveis, como por exemplo a variável **X**. Os valores individuais que uma variável pode assumir são representados pelas correspondentes letras minúsculas. Por exemplo se **X** é usado para designar o peso de uma amostra de 50 pessoas, então **x** é o valor numérico do peso de uma dessas 50 pessoas. Diferentes valores de uma variável são identificados por subscritos. Assim, os pesos de 50 pessoas em uma amostra podem ser denotados por $x_1, x_2, \ldots, x_{50}$.

- número total de observações em uma população finita é designado por **N** e na amostra é representado por **n**. A distinção entre medidas descritivas para populações e amostras é muito importante. Denotaremos os parâmetros (medidas referentes a população) por letras gregas ou letras minúsculas em português. As estatísticas amostrais serão representadas por letras maiúsculas em português e os valores observados de uma estatística amostral pela correspondente letra minúscula em português. Por exemplo, as medidas descritivas a serem introduzidas nessa seção serão denotadas como segue:

2.5.2 A Média Aritmética Não Ponderada

A média é definida como a soma das observações dividida pelo número de observações. Se tivermos, por exemplo, **n** valores, temos:

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

Propriedades da média aritmética não ponderada:

1. A média é um valor típico, ou seja, ela é o centro de gravidade da distribuição, um ponto de equilíbrio. Seu valor pode ser substituído pelo valor de cada item na série de dados sem mudar o total. Simbolicamente temos:

$$n(\overline{X}) = \sum x \qquad (^5)$$

2. A soma dos desvios das observações em relação a média é igual a zero.

$$\sum (x - \overline{X}) = 0$$

3. A soma dos desvios elevados ao quadrado das observações em relação a média é menor que qualquer soma de quadrados de desvios em relação a qualquer outro número. Em outras palavras,

$\sum (x - \overline{X})^2$ = é um mínimo.

(1) x	(2) $\overline{x}$	(3) $(x - \overline{x})$	(4) $(x - \overline{x})^2$	(5) $(x - 2)^2$	(6) $(x - 5)^2$
1	3	-2	4	1	16
2	3	-1	1	0	9
6	3	+3	9	16	1
Soma	9	0	14	17	26

2.5.3 A Média Aritmética Ponderada

No cálculo da média aritmética não ponderada todos os valores observados foram somados atribuindo-se o mesmo peso a todas observações. Agora veremos uma nova forma de calcular a média. Consideremos um exemplo familiar de cálculo da média de notas de estudantes, quando o exame final vale duas vezes mais do que as duas provas comuns realizadas no decorrer do semestre. Se um determinado obter as notas 7, 5 e 8 a sua média ponderada final será:

$$\frac{1\times(7) + 1\times(5) + 2\times 8}{1+1+2} = 7$$

Em termos gerais, a fórmula para a média aritmética ponderada é:

$$\overline{X}_w = \sum_{i=1}^{n} w_i \times x_i = \sum wx$$

onde $\mathbf{w_i}$ é o peso da observação **i**

e **n** é o número de observações.

2.5.4 Proporções como Médias

Suponha que queremos determinar a proporção de votantes entre os cidadãos brasileiros. Devemos primeiro designar um valor 1 para cada pessoa qualificada como eleitor e um valor 0 para cada pessoa não qualificada como eleitor. Então, a soma dos 1's seria $\sum x$ e a média seria a média seria obtida pela divisão da soma pelo número **N** total de pessoas no Brasil.

---

[5] - Utilizaremos muito freqüentemente a notação $\sum x$ simplificadamente para representar $\sum_{i=1}^{n} x_i$.

A média da variável **x** é $\mu = \sum x / N$. No entanto essa média é também uma proporção, a proporção de eleitores na população brasileira.

2.5.5 A Média Geométrica

A média geométrica de uma amostra é definida como a raiz enésima do produto nos **n** valores amostrais.

$$G = \sqrt[n]{(x_1)(x_2)...(x_n)}$$

Por exemplo, a média geométrica de 5, 9 e 13 é:

$$G = \sqrt[3]{(5)(9)(13)} = 8{,}36$$

Para a mesma série de dados a média é 9. É sempre verdade que a média aritmética é maior do que a média geométrica para qualquer série de valores positivos, com exceção do caso em que os valores da série são todos iguais, quando as duas médias coincidem.

A conclusão que chegamos é que o logaritmo da média geométrica é igual a média aritmética dos logaritmos dos valores da série. Verifica-se que a média geométrica somente tem significado quando todos os valores da série são todos positivos.

Suponhamos como exemplo de aplicação de cálculo da média geométrica os dados da tabela seguinte que mostram as mudanças de preços de duas mercadorias, **A** e **B**, de 1980 a 1985. Durante esse período o preço de **A** subiu 100 % e o preço de **B** decresceu 50 %. Qual foi a mudança média relativa de preços? Em outras palavras, qual foi o percentual médio de mudança de preços?

A mais importante aplicação da média geométrica refere-se talvez ao cálculo de taxas de crescimento médias, desde que essas podem ser corretamente medidas somente por esse método. Para exemplificar, no campo da economia, esse ponto, suponha que a produção anual de um setor industrial cresceu de 10.000 para 17.280 unidades durante o período 1985-1988 como mostrado na tabela a seguir; qual é a taxa média de crescimento anual? A taxa média anual de crescimento pode ser calculada a partir dos valores em porcentagem da produção em relação aos anos anteriores. Se calcularmos a média aritmética desses valores teríamos:

$$\bar{x} = (60 + 96 + 300)/3 = 152$$

implicando uma taxa de crescimento média de 152 -100 = 52 %. Se a produção cresce 52 % ao ano, começando da produção de 1985 de 10.000 unidades, então a produção de 1986 seria de

23.0 + 0,52 (10.000) = 15.200;

a produção de 1987 seria de

15.200+ 0,52(15.200) = 23.104;

a produção de 1988 seria de

23.104 + 0,52(23.104) = 35.118,08

Ano	1985	1986	1987	1988
Produção	10.000	6.000	5.760	17.280
Porcentagem do ano anterior	-	60	96	300

Observe-se que este último valor é quase 200 % do valor efetivamente observado em 1988, de 17.200.

A média geométrica, por sua vez, é: $g = \sqrt[3]{(60)(96)(300)} = 120$

implicando uma taxa anual média de crescimento de 120 - 100 = 20 %. Verificando, teremos:

no ano de 1986: 10.000 + 0,20(10.000) = 12.000;

no ano de 1987: 12.000 + 0,20(12.000) = 14.400;

no ano de 1988: 12.000 + 0,20(14.400) = 17.280 que coincide com o valor observado efetivamente em 1988.

Se o valor da média geométrica das porcentagens de crescimento for menor do que 100, implica em uma porcentagem média de crescimento negativa, o que indica uma taxa média de declínio ao invés de uma taxa média de crescimento.[6] Atente também para o fato de que as três porcentagens a partir das quais a média geométrica é calculada são *percentuais do ano anterior* ao invés de *mudança percentual do ano anterior.*[7]

- cálculo da taxa média de crescimento é baseado principalmente na hipótese de uma taxa constante de crescimento ou de que os valores individuais formam uma progressão geométrica. Quando o cálculo envolve um número considerável de períodos, utiliza-se com mais freqüência uma fórmula que se relaciona com a média geométrica, que é:

$$R = \left(\sqrt[n]{\frac{x_f}{x_i}} \right) - 1$$

onde:

R = taxa de crescimento geométrica média,

n = número de períodos de tempo,

x_f = valor no período final,

x_i = valor no período inicial.

2.5.6 A Média Harmônica

A média harmônica é o inverso da média aritmética dos inversos dos valores observados. Simbolicamente, para uma amostra, temos:

$$H = \frac{\dfrac{1}{1/x_1 + 1/x_2 + \ldots + 1/x_n}}{n} = \frac{\dfrac{1}{\sum(1/x)}}{n} = \frac{n}{\sum(1/x)}$$

Para cálculos mais simples, a fórmula anterior pode ser reescrita como:

---

[6] Se, por exemplo, ao invés de 60, 96 e 300 %, como anteriormente, tivermos 60, 96 e 78 %, a taxa de crescimento geométrica média será de $g = \sqrt[3]{(60)(96)(78)} = 76{,}59$, o que indica um decréscimo médio de 76,59 - 100 = - 23,41 %.

[7] Essas últimas porcentagens, referentes ao exemplo da tabela anterior, seriam (6.000 - 10.000)/10.000= - 0,40, ou seja - 40 %; (5.760 - 6.000)/6.000 = -0,04 , ou seja, - 4 %; e (17.280 - 5.760)/5760 = 2, ou seja + 200 %.

$$1/H = \frac{1/x_1 + 1/x_2 + ... + 1/x_n}{n} = \frac{\sum(1/x)}{n}$$

Calcule a média harmônica dos três valores 4, 10 e 16 :

Para os mesmos dados a média aritmética é 10 e a média geométrica é 8,62. Para qualquer série de dados cujos valores não são todos os mesmos e que não incluem o zero, a média harmônica é sempre menor que tanto a média aritmética como a média geométrica.

2.5.7 A Mediana

A mediana é o valor do item central da série quando estes são arranjados em ordem de magnitude. Para a série R$ 2, R$ 4, R$ 5, R$ 7 e R$ 8, a mediana é o valor do terceiro item, R$ 5. No caso do número de itens na série ser par, a mediana é a semi-soma dos dois valores mais centrais. Por exemplo, para a série 3, 5 ,8 ,10, 15 e 21 kg, a mediana é a media dos valores 8 e 10, ou seja 9.

A mediana pode ser formalmente definida como o valor que divide a série de tal forma que no mínimo 50 % dos itens são iguais ou menores do que ela, e no mínimo 50 % dos itens são iguais ou maiores do que ela. Mais rigorosamente, estabelecemos que:

$X_{.5}$ = o valor do [(n+1)/2] -ésimo item

Por exemplo, para uma série formada pelos valores 3,5,8,10,15 e 21 a mediana será o valor do [(6+1)/2] = 3,5 ésimo item, ou seja, a semi soma do item de posto 3 e do item de posto 4, que são 8 e 10.

O valor da mediana não é influenciado pelos valores nas caudas de uma distribuição. Por exemplo, se temos a série de dados 1,2,3,4,5 a mediana é 3. Se substituímos os valores das caudas dessa distribuição por quaisquer valores uma nova distribuição formada poderia ser formada pela série -1000,-100,3,500,5000 e a mediana permanece sendo 3. Portanto, ela é uma medida de posição da distribuição bem adequada para distribuições assimétricas, tais como a distribuição de renda, já que não sabemos se a família mais rica ganha R$7.000.000 ou R$ 500.000.000. Veremos, mais a frente que ela possui vantagens em relação a média aritmética, como medida de posição (ou medida de tendência central) para dados agrupados em classes de freqüência, quando a última classe tem limite superior indeterminado.

A mediana também tem a interessante propriedade de que a soma dos desvios absolutos das observações em relação a mediana é menor do que a soma dos desvios absolutos a partir de qualquer outro ponto na distribuição. Simbolicamente:

$$\sum|x - X_{.5}| = \text{um mínimo}$$

2.5.8 A Média para Dados Agrupados

Quando estamos tratando de amostras ou populações muito grandes é conveniente calcular as medidas descritivas a partir das distribuições de freqüência. A média não pode ser determinada exatamente a partir de distribuições de freqüência, mas uma boa aproximação pode ser obtida pela hipótese do ponto médio. A aproximação é quase sempre muito satisfatória se a distribuição é bem

construída.[8] A hipótese do ponto médio refere-se a considerar-se de que todas as observações de uma dada classe estão centradas no ponto médio daquela classe. Conseqüentemente, o valor total da freqüência da classe da i-ésima classe é simplesmente o produto $\mathbf{f_i}\ \mathbf{m_i}$, onde $\mathbf{f_i}$ é a freqüência (absoluta simples) da classe **i** e $\mathbf{m_i}$ é ponto médio da classe **i**. Sob essa hipótese, a média aproximada para uma distribuição de uma amostra com **k** classes vem a ser:

$$\overline{X} \cong \frac{f_1 m_1 + f_2 m_2 + ... + f_k m_k}{f_1 + f_2 + ... + f_k} \cong \frac{\sum fm}{\sum f}$$

$$= \frac{\sum fm}{n}$$

É importante notar que todos os somatórios na equação acima referem-se às classes e não às observações individuais. Consideremos a seguinte tabela de distribuição de freqüência para dados de gasto com alimentação extraídos de uma pesquisa de orçamentos familiares.

Classe	f_i	M	fm
R$ 120,00 - 139,99	5	130,0	650,0
R$ 140,00 - 159,99	26	150,0	3900,0
R$ 160,00 - 179,99	24	170,0	4080,0
R$ 180,00 - 199,99	15	190,0	2850,0
R$ 200,00 - 219,99	8	210,0	1680,0
R$ 220,00 - 239,99	2	230,0	460,0
Total	80		13620,0

$$\overline{x} = \frac{13620,00}{80} = R\$170,25$$

Ao utilizar essa aproximação estamos considerando a hipótese de que todas as observações em cada classe estão uniformemente distribuídas nessa classe. Por exemplo, se tivermos um intervalo de tamanho 100 e com freqüência igual a 6 observações, a localização dessas observações seria 0,20,40,60,80 e 100, com distância constante entre cada par de observações, de forma que:
0+20+40+60+80+100 = 300 = m x 6 e m = 50 , ou seja o ponto médio do intervalo de 0 a 100. Conclui-se que se a distribuição das observações for uniforme em cada intervalo, o somatório dos valores das observações de cada intervalo é igual ao produto da freqüência no intervalo pelo valor do ponto médio desse intervalo. Supõe-se que com uma conveniente construção de intervalos de classe os eventuais erros nos intervalos compensam-se mutuamente.

2.5.9 A Mediana para dados Agrupados

Assim como é possível estabelecer uma aproximação da média aritmética para dados agrupados, o mesmo pode ser feito para a mediana. O método usado é o da interpolação utilizando-se a distribuição de freqüência acumulada ou ogiva. Inicialmente determina-se a classe que contem a mediana. Essa será a classe cuja freqüência acumulada relativa correspondente a seu limite inferior é

[8] Isto é, principalmente se no agrupamento dos dados originais em uma tabela de distribuição de freqüência, empregou-se um número adequado de classes de freqüência.

menor que 0,50 (ou 50 %) e a freqüência acumulada relativa correspondente a seu limite superior é maior que 0,50 (ou 50 %). O próximo passo é a determinação do ponto exato onde localiza-se a mediana naquela classe. Para o exemplo anterior de gastos com alimentação de famílias, temos:

Classe	freq. absoluta simples	freq.acumulada	freqüência relativa acumulada
R$ 120,00 - R$139,99	5	5	0,0625
140,00 - 159,99	26	31	0,3875
160,00 - 179,99	24	55	0,6875
180,00 - 199,99	15	70	0,8750
200,00 - 219,99	8	78	0,9750
220,00 - 239,99	2	80	1,0000
Total	80		

A classe que contém a mediana é a terceira classe, pois a freqüência relativa acumulada da classe anterior (segunda classe) é menor que 0,5 e a freqüência relativa acumulada da terceira classe é maior do que 0,5.[9]

Freqüência acumulada da classe que contem a mediana

c = amplitude (tamanho) da classe da mediana.

2.5.10 A Moda para dados Agrupados

A moda de uma distribuição de freqüência pode muitas vezes ser aproximada pelo ponto médio da classe modal - a classe com maior densidade de freqüência.[10] Então, para os dados de gastos com alimentação do exemplo anterior, x_m = R$ 150, o ponto médio da segunda classe, que possui a maior freqüência. Esse método de localizar a moda é totalmente satisfatório quando as densidades de freqüência da classe imediatamente anterior à classe modal (a classe premodal) e da classe imediatamente posterior à classe modal (classe posmodal) são aproximadamente iguais. Quando isso não ocorre, como sugerido pela figura a seguir, resultados mais precisos podem ser obtidos com a seguinte fórmula, para uma amostra:

$$X_m \cong L_m + (\frac{\Delta_1}{\Delta_1 + \Delta_2})c \quad \text{ou } M_o = l_i + \left(\frac{d_1}{d_1 + d_2}\right)h$$

onde:

L_m = o verdadeiro[11] limite inferior de classe da classe modal

Δ_1 = da diferença entre das densidades de freqüência da classe modal e classe premodal.

Δ_2 = da diferença entre das densidades de freqüência da classe modal e classe posmodal.

C = a verdadeira amplitude de classe da classe modal.

9 -

10

11

No exemplo anterior de gastos com alimentos de 80 famílias, como a amplitude de todos os intervalos são iguais, podemos utilizar as freqüências absolutas de classe no lugar das densidades de freqüência, para o cálculo do valor aproximado da mediana.

$\mathbf{L_m} = 140{,}00$ $\quad\quad$ $\mathbf{\Delta_1} = 26 - 15 = 11$

$\mathbf{c} = 20$ $\quad\quad$ $\mathbf{\Delta_2} = 26 - 24 = 2$

$$x_m \cong 140{,}00 + (\frac{11}{11+2})20 = 156{,}92$$

Uma observação é aqui necessária. É possível calcular os valores aproximados da mediana e da moda para dados agrupados quando o último intervalo de classe tem limite superior indeterminado. No caso da mediana isso é imediato e no caso da moda, o seu cálculo somente pode ser feito se a última classe não for a classe modal e é preciso primeiramente calcular as densidades de freqüência. Como exemplo, suponhamos que a distribuição de renda de uma certa região é dada pela seguinte distribuição de freqüência:

renda (R$) limites nominais	limites reais	freqüência absoluta	densidade de freqüência
0 - 120	0 - 120,50	40	40/120,50 = 0,332
121 - 605	120,50 - 605,50	170	170/485 = 0,350
606 - 1200	605,50 - 1200,50	220	220/595 = 0,370
1201 - 2400	1250,50 - 2400,50	15	15/1150 = 0,013
Mais de 2400	mais de 2450,50	97	indeterminado
Total		542	

A mediana está localizada na terceira classe:[12]

$$x_{.5} \cong 605{,}50 + \left[\frac{(542+1)/2 - 210}{220}\right](1200{,}50 - 605{,}50) = 772$$

A classe modal também é a terceira classe:[13]

$$x_m = 605{,}50 + \frac{(0{,}370 - 0{,}350)}{(0{,}370 - 0{,}350) + (0{,}370 - 0{,}013)}(1200{,}50 - 605{,}50) = 637$$

Infelizmente, para esse exemplo não é possível o cálculo da média, o que demonstra que para algumas situações temos que contar com a mediana como medida de posição (ou de tendência central) de uma distribuição estatística.

Discutiremos agora comparativamente algumas das características das três principais medidas de posição:

A Média Aritmética

1) Ela é afetada por todas as observações e é influenciada pelas magnitudes absolutas dos valores extremos na série de dados.

13

2) Ela é das três medidas de posição a que possibilita maiores manipulações algébricas, dadas as características de sua fórmula.

3) Em amostragem, a média é uma estatística estável. Isso será aprofundado posteriormente.

A Mediana

1) Seu valor é afetado pelo número de observações e como elas estão distribuídas mas ela não é afetada pelos valores das observações extremas.

2) Sua fórmula não é passível de manipulação algébrica.

3) Seu valor pode ser obtido, como vimos, em distribuições, com limites superiores indeterminados para a sua última classe.

4) A mediana é a estatística mais adequada para descrever observações que são ordenadas ao invés de medidas.

A Moda

1) A moda é o valor mais típico e representativo de uma distribuição. Ela representa o seu valor mais provável.

2) Como a mediana, a moda também não é influenciada pelos valores extremos da distribuição e não permite manipulações algébricas como a fórmula da média.

Existem algumas relações entre as diversas medidas de posição:

1) Para qualquer série, exceto quando no caso de todas as observações coincidirem em um único valor, a média aritmética é sempre maior que a média geométrica, a qual, por sua vez, é maior que a média harmônica.

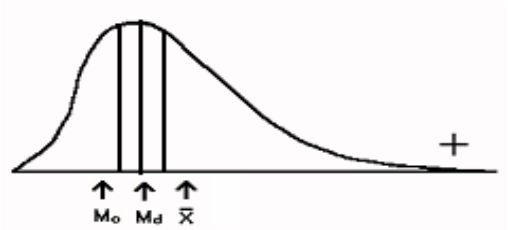
2) Para uma distribuição simétrica e unimodal, **média = mediana = moda.**

3) Para uma distribuição positivamente assimétrica, **média > mediana > moda.** A distância entre a mediana e a média é cerca de um terço da distância entre a moda e a média.

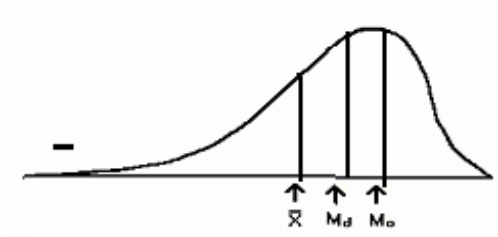
4) Para uma distribuição negativamente assimétrica, **média < mediana < moda**. A distância entre a mediana e a média é cerca de um terço da distância entre a moda e a média.Essas últimas características são apresentadas graficamente, a seguir

POSIÇÕES RELATIVAS DA MÉDIA, MEDIANA E MODA EM FUNÇÃO DA ASSIMETRIA DAS DISTRIBUIÇÕES

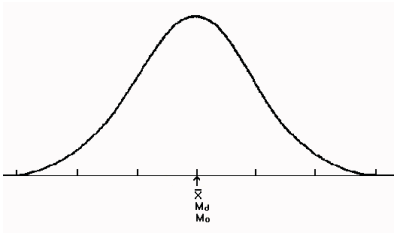
Assimetria positiva



Assimetria negativa



Distribuição Simétrica



Medidas de Dispersão, Assimetria e Curtose

Muitas séries estatísticas podem apresentar a mesma média, mas no entanto, os dados de cada uma dessas séries podem distribuir-se de forma distinta em torno de cada uma das médias dessas séries. Na análise descritiva de uma distribuição estatística é fundamental, além da determinação de uma medida de tendência central, conhecer a dispersão dos dados e a forma da distribuição. Duas séries de dados podem possuir a mesma média, mas uma pode apresentar valores mais homogêneos (menos dispersos em relação a média) do que a outra. Um país, por exemplo, com uma distribuição de renda mais equânime, terá uma dispersão de suas rendas menor do que um país com estrutura de renda mais diferenciada em diversos estratos ou categorias sociais. Uma máquina que produz parafusos e que estiver menos ajustada do que outra produzirá medidas de parafusos com distribuição mais dispersa em torno de sua média.

A inequação das médias

A importância das médias é com freqüência exagerada. Se dizemos que a renda familiar média de um determinado país é de US$ 5.000 por ano não sabemos muita coisa sobre a distribuição de renda desse país. Uma média, como um simples valor adotado para representar a tendência central de uma série de dados é uma medida muito útil. Porém, o uso de um simples e único valor para descrever uma distribuição abstrai-se de muitos aspectos importantes.

Em primeiro lugar, nem todas as observações de uma série de dados tem o mesmo valor da média. Quase sem exceção, as observações incluídas em uma distribuição distanciam-se do valor central, embora o grau de afastamento varie de uma série para outra. Muito pouco pode ser dito a respeito da dispersão mesmo quando diversas medidas de tendência central são calculadas para a série. Por exemplo, não podemos dizer qual distribuição tem maior ou menor grau de dispersão da informação dada pela tabela abaixo.

	Distribuição A	Distribuição B
Média	15	15
Mediana	15	12
Moda	15	6

Uma segunda consideração é que as formas de distribuição diferem de um conjunto de dados para outro. Algumas são simétricas; outras não. Assim, para descrever uma distribuição precisamos também de uma medida do grau de simetria ou assimetria. A estatística descritiva para esta característica é chamada de *medida de assimetria*.

Finalmente, existem diferenças no grau de achatamento entre as diferentes distribuições. Esta propriedade é chamada de *curtose* (em inglês, *kurtosis*). Medir a curtose de uma distribuição significa comparar a concentração de observações próximas do valor central com a concentração de observações próximas das extremidades da distribuição.

2.5.11 O Intervalo (ou amplitude)

A medida de dispersão mais simples é a *amplitude*, a diferença entre o maior e o menor valor nos dados. Para uma distribuição de freqüência que usa intervalos de classe, a amplitude pode ser considerada como a diferença entre o maior e o menor limite de classe ou a diferença entre os pontos médios dos intervalos de classe extremos. Os preços de ações e de outros ativos financeiros são freqüentemente descritos em termos de sua amplitude, com a apresentação pelas Bolsas de Valores do maior valor e do menor valor da ação em um determinado período de tempo.

Para algumas distribuições simétricas a média pode ser aproximada tomando-se a semi-soma dos dois valores extremos,[14] que é freqüentemente chamada de semi-amplitude. Por exemplo, é prática entre os meteorologistas derivar a média diária de temperatura tomando a média somente dos valores máximo e mínimo de temperatura ao invés, de digamos, a média das 24 leituras horárias do dia.

A amplitude tem alguns defeitos sérios. Ela pode ser influenciada por um valor atípico na amostra. Além disso o seu valor independe do que ocorre no interior da distribuição, já que somente depende dos valores extremos.

2.5.12 Percentis, Decis e Quartis

Podemos tentar responder a seguinte pergunta: “que proporção dos valores de uma variável é menor ou igual a um dado valor? Ou maior ou igual a um dado valor? Ou entre dois valores ?” Quando construímos uma distribuição de freqüência acumulada, tais questões somente podem ser respondidas com relação aos limites de classe exatos. Por exemplo, a partir da distribuição de freqüência relativa acumulada da página 28, podemos dizer que 38,75 % das observações são menores do que 159,99. Mas não podemos responder a pergunta: “qual é o gasto familiar tal que a proporção da amostra tendo este valor ou menos é 35 %?”. Mas é visível da tabela que 6,25 % das famílias gastam com alimentação até R$ 139,99 e 38,75 % das famílias gastam até R$ 159,99. Portanto, como 35 % está entre estes dois valores, o gasto familiar tal que a proporção da amostra tendo este valor ou menos é 35 % está situado entre R$ 139,99 e R$ 159,99. Este valor é chamado de percentil 35.

O percentil 40 é o valor da variável que é maior do que 40 % das observações. Generalizando, o percentil **x**, é o valor da variável que é maior do que **x** % das observações. Em outras palavras, o percentil **x** é o valor da variável correspondente ao valor de freqüência relativa acumulada de **x** %.[15] O

---

[15] Para o cálculo do valor exato do percentil **x** para dados agrupados utiliza-se o mesmo método para a determinação da mediana, ou seja, a interpolação linear. Como no caso da mediana, podemos empregar uma fórmula de interpolação $X_p = LI_p + \left[\frac{p\times(n+1)/100 - F_a}{f_p}\right]c$

onde $\mathbf{X_p}$ é o percentil **p**, $\mathbf{Li_p}$ é o limite inferior real da classe que contem o percentil, $\mathbf{F_a}$ é a freqüência relativa acumulada da classe anterior à classe que contem o percentil, $\mathbf{f_p}$ é a freqüência relativa (simples)

primeiro decil é o valor da variável que supera um décimo (ou 10 %) do total de observações. Se tivermos 200 observações, o segundo decil será aproximadamente a observação de posto 40.

O primeiro quartil é o valor da variável cuja freqüência relativa acumulada é 0,25 (ou 25 %). O terceiro quartil é o valor da variável cuja freqüência relativa acumulada é 0,75 (ou 75 %). O primeiro quartil é maior do que um quarto dos valores observados e menor do que três quartos destes valores. O terceiro quartil é maior do que três quartos dos valores observados e menor do que um quarto destes valores. O segundo quartil confunde-se com a mediana.

Uma medida de dispersão é o chamado desvio interquartílico que é a diferença entre o terceiro e o primeiro quartis.

2.5.13 Variância e Desvio Padrão

A variância é definida como a média dos desvios ao quadrado em relação à média da distribuição. Para uma amostra,

$$S^2 = \frac{\sum (x - \overline{X})^2}{n-1}$$

Para uma população finita,

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Na penúltima equação, **n-1** é chamado de número de "graus de liberdade" de $\mathbf{S^2}$, um conceito a ser definido mais tarde. Existe uma restrição para esta equação: **n > 1** (não se pode calcular a variância para uma amostra de uma observação apenas). O desvio padrão é a raiz quadrada da variância, e é denotado **S** (para amostra) e **σ** (para população). Existem fórmulas que facilitam os cálculos para $\mathbf{S^2}$ e $\boldsymbol{\sigma^2}$ **:**

$$S^2 = \frac{n\sum x^2 - (\sum x)^2}{n(n-1)}$$

$$\sigma^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$$

Com estas duas últimas fórmulas, podemos calcular a variância somente com a soma dos valores ($\sum x$) e a soma dos quadrados dos valores ($\sum x^2$); não é mais necessário calcular a média, em seguida os desvios em relação às médias e finalmente os quadrados destes desvios.

Para ilustrar o processo de cálculo da variância e desvio padrão e para mostrar o uso destas medidas, considere o seguinte exemplo. Dois tipos diferentes de máquina, **X** e **Y** são projetadas para produzir o mesmo produto. Elas têm o mesmo preço de venda. Um fabricante está tentando decidir qual delas comprar e observou 10 máquinas distintas de cada tipo em operação por uma hora. A tabela seguinte mostra as produções horárias nas primeiras duas colunas. As médias são $\overline{x} = {}^{403}\!/_{10} = 40{,}3$ unidades por hora e $\overline{y} = {}^{408}\!/_{10} = 40{,}8$ unidades por hora. Portanto, com base nestes dados, o tipo **Y** é

---

da classe que contem o percentil, **c** é a amplitude do intervalo de classe que contem o percentil e **é** o número de observações. O mesmo método pode ser empregado também para os decis e quartis.

um pouco mais rápida. Podemos retirar mais alguma informação a partir destes dados? Podemos medir e comparar as dispersões das produções horárias dos dois tipos de máquina. Usando a penúltima fórmula para os dados da tabela, obtemos:

$$S_X^2 = \frac{10(16.405) - (403)^2}{10(10-1)} = 18,23$$

$$S_X = \sqrt{18,23} = 4,27 \text{ unidades por hora}$$

$$S_Y^2 = \frac{10(17.984) - (408)^2}{10(10-1)} = 135,11$$

$$S_Y = \sqrt{135,12} = 11,62 \text{ unidades por hora}$$

x	Y	x^2	Y^2
35	25	1.225	625
36	26	1.296	676
49	55	2.401	3.025
44	52	1.936	2.704
43	48	1.849	2.304
37	24	1.369	576
38	34	1.444	1.156
42	47	1.764	2.209
39	50	1.521	2.500
40	47	1.600	2.209
403	408	16.405	17.984

O tipo **X** tem menor dispersão que o tipo **Y.** Apesar de ter maior preço que o tipo **Y,** a máquina **X** é mais precisa.

2.5.14 Variância e Desvio Padrão para Dados Agrupados

A variância e o desvio padrão (como a média, mediana, moda, quartis, percentis, decis) podem ser calculados para dados agrupados, ou seja, distribuições de freqüência com intervalos de classe. Entretanto, os resultados podem ser apenas aproximadamente precisos. Utiliza-se, como no caso da média, a hipótese do ponto médio: a de que toda observação está localizada no ponto médio de sua classe. Cada ponto médio entra nos cálculos quantas vezes são as observações naquele intervalo de classe. As equações para as variâncias são:

$$S^2 = \frac{\sum f(m - \overline{X})^2}{n-1}, \text{ para a amostra;}$$

$$\sigma^2 = \frac{\sum f(m - \mu)}{N}, \text{ para a populacao.}$$

Os símbolos utilizados nestas equações já foram definidos anteriormente. Para facilitar os cálculos podemos utilizar as seguintes fórmulas mais convenientes para as variâncias:

$$S^2 = \frac{\sum fm^2 - (\sum fm)^2 / n}{n-1}$$

e

$$\sigma^2 = \frac{\sum fm^2 - (\sum fm)^2 / N}{N}$$

para a amostra e população, respectivamente. Aqui, como antes, assumimos que a população é finita.

Os somatórios em todas estas equações são para todas as k classes, não para as observações individuais. Estas equações podem ser aplicadas tanto para intervalos de classe iguais como para intervalos de classe desiguais. Entretanto, elas não podem ser empregadas quando existem um ou mais intervalos sem limites. Como para os dados não agrupados, a raiz quadrada destas equações são os desvios padrões para a amostra e para a população, respectivamente.

Aplicando as últimas equações para o exemplo de consumo de alimentos, temos:

Classe	(1) m	(2) f_i	(3) fm (2)(1)	(4) fm^2 (3)(1)
R$ 120,00 - R$139,99	130	5	650	84.500
140,00 - 159,99	150	26	3.900	585.000
160,00 - 179,99	170	24	4.080	693.000
180,00 - 199,99	190	15	2.850	541.500
200,00 - 219,99	210	8	1.680	352.800
220,00 - 239,99	230	2	460	105.800
Total		80	13.620	2.363.200

$$S^2 = \frac{\sum fm^2 - (\sum fm)^2 / n}{n-1} = \frac{2.363.200 - (13.620)^2 / 80}{80-1} = 561{,}96$$

$$S = \sqrt{561{,}96} = 23{,}71$$

2.5.15 Interpretando e Aplicando o Desvio Padrão

O desvio padrão é mais a mais usada das medidas de variabilidade. Infelizmente, o desvio padrão não tem uma interpretação intuitivamente óbvia. Por exemplo, no exemplo anterior das máquinas, S_X = 4,27 unidades por hora, mas não é óbvio o que isto quer dizer para a máquina X. Para muitas séries de dados há dois teoremas para a interpretação do desvio padrão que são muito úteis. Eles são chamados de Desigualdade de Chebyshev e a Regra de Gauss, as quais introduzimos a seguir.

Teorema: A Regra de Gauss. Se os dados são amostrais e se são, de forma aproximada, distribuídos normalmente, ou seja, o histograma dos dados é aproximadamente simétrico e tem a forma de um sino, então:

1. $\overline{X} \pm 1S$ incluirá aproximadamente 68 % dos dados
2. $\overline{X} \pm 2S$ incluirá aproximadamente 95 % dos dados
3. $\overline{X} \pm 3S$ incluirá aproximadamente 100 % dos dados

Chamamos isto de Regra de Gauss, porque é baseada na distribuição de probabilidade gaussiana (ou distribuição de probabilidade normal). Esta distribuição será discutida em detalhe em um capítulo posterior.

2.5.16 Coeficiente de Variação

Com freqüência, como no caso do exemplo das duas máquinas, queremos comparar a variabilidade de dois ou mais conjuntos de dados. Podemos fazer isto facilmente usando as variâncias ou os desvios padrões quando, primeiro, todas as observações individuais têm a mesma unidade de medida e, segundo, as médias dos conjuntos de dados são aproximadamente iguais. Quando qualquer uma destas condições não é satisfeita, uma medida relativa de dispersão deve ser usada. Uma medida relativa de variabilidade freqüentemente usada é chamada de *coeficiente de variação,* denotada por CV para uma amostra. Esta medida é o valor do desvio padrão em relação à média:

$$CV = \frac{S}{\overline{X}}$$

Suponha que um cientista na Índia obteve os seguintes dados referentes aos pesos de elefantes e ratos.

Elefantes	Ratos
$\overline{x}_E = 6.000 \text{ kg}$	$\overline{x}_R = 0{,}150 \text{ kg}$
$s_E = 300 \text{ kg}$	$s_R = 0{,}04 \text{ kg}$

Se calcularmos os respectivos coeficientes de variação, teremos:

$$cv(X_E) = \frac{s_E}{\overline{x}_E} = \frac{300}{6000} = 0{,}050 \quad \text{ou } 5{,}0\,\%$$

$$cv(X_R) = \frac{s_R}{\overline{x}_R} = \frac{0{,}04}{0{,}150} = 0{,}266 \quad \text{ou } 26{,}7\,\%$$

Portanto, a variabilidade relativa dos pesos dos ratos é mais do que 5 vezes maior do que a variabilidade dos pesos dos elefantes. Para o exemplo anterior das máquinas, teremos:

$$cv(X) = \frac{4{,}27}{40{,}30} = 0{,}1060 \quad \text{ou } 10{,}60\,\%$$

$$cv(Y) = \frac{11{,}62}{40{,}80} = 0{,}2848 \quad \text{ou } 28{,}48\,\%$$

Assim, a dispersão relativa da produção da máquina Y é quase três vezes maior do que a dispersão relativa da máquina X.

2.6 Medidas de Assimetria

Duas distribuições também podem diferir uma da outra em termos de assimetria ou achatamento, ou ambas. Como veremos, assimetria e achatamento (o nome técnico utilizado para esta última característica de forma da distribuição é *curtose*) têm importância devido a considerações teóricas relativas à inferência estatística que são freqüentemente baseadas na hipótese de populações distribuídas normalmente. Medidas de assimetria e de curtose são, portanto, úteis para se precaver contra erros aos estabelecer esta hipótese.

Diversas medidas de assimetria são disponíveis, mas introduziremos apenas uma, que oferece simplicidade no conceito assim como no cálculo. Esta medida, a *medida de assimetria de Pearson*, é baseada nas relações entre a média, mediana e moda. Recorde que estas três medidas são idênticas em valor para uma distribuição unimodal simétrica, mas para uma distribuição assimétrica a média distancia-se da moda, situando-se a mediana em uma posição intermediária, a medida que aumenta a assimetria da distribuição. Conseqüentemente, a distância entre a média e a moda poderia ser usada para medir a assimetria. Precisamente,

Assimetria = média - moda

Quanto maior é a distância, seja negativa ou positiva, maior é a assimetria da distribuição. Tal medida, entretanto, tem dois defeitos na aplicação. Primeiro, porque ela é uma medida absoluta, o resultado é expresso em termos da unidade original de medida da distribuição e, portanto, ela muda quando a unidade de medida muda. Segundo, a mesma grandeza absoluta de assimetria tem diferentes significados para diferentes séries de dados com diferentes graus de variabilidade. Para eliminar estes defeitos, podemos medir uma medida relativa de assimetria. Esta é obtida pelo *coeficiente de assimetria de Pearson*, denotado por $\mathbf{SK_P}$ e dado por:

$$SK_P = \frac{\overline{X} - X_m}{S}$$

A aplicação desta expressão envolve outra dificuldade, que surge devido ao fato de que o valor modal da maioria das distribuições ser somente uma distribuição, enquanto que a localização da mediana é mais satisfatoriamente precisa. Contudo, em distribuições moderadamente assimétricas, a expressão

$$X_m = \overline{X} - 3(\overline{X} - X_{.5})$$

é adequada (não envolve imprecisão muito grande). A partir disto, vemos que:

$$\overline{X} - X_m = \overline{X} - [\overline{X} - 3(\overline{X} - X_{.5})] = 3(\overline{X} - X_{.5})$$

Com este resultado, podemos rescrever o coeficiente de assimetria de Pearson como:

$$SK_P = \frac{3(\overline{X} - X_{.5})}{S}$$

Esta medida é igual a zero para uma distribuição simétrica, negativa para distribuições com assimetria para a direita e positiva para distribuições com assimetria para a esquerda. Ela varia dentro dos limites de ± 3. Aplicando $\mathbf{SK_P}$ aos dados agrupados de gastos com consumo de alimentos das famílias, temos:

$$SK_P = \frac{3(170{,}25 - 167{,}92)}{23{,}71} = +0{,}295$$

Este resultado revela que a distribuição de gastos com consumo de alimentos tem assimetria moderadamente positiva (o que significa maior concentração de famílias nas classes de menor gasto). É muito comum encontrar distribuições positivamente assimétricas em dados econômicos, particularmente na produção e séries de preços, os quais podem ser tão pequenos quanto nulos mas podem ser infinitamente grandes. Distribuições assimetricamente negativas são raras em ciências sociais.