Resumo 3 – Resumo dos dados numéricos por meio de números

1. Medidas de Tendência Central

A *tendência central* da distribuição de freqüências de uma variável em um conjunto de dados é caracterizada pelo *valor típico* dessa variável. Essa é uma maneira de resumir a informação contida nos dados, pois escolheremos um valor para representar todos os outros.

1.1. Média Aritmética Simples ($\overline{X}$)

A média aritmética, ou simplesmente média, é, sem dúvida, a medida de posição mais utilizada.

O símbolo μ (mi) é usado para denotar a média de uma população e $\overline{X}$ (x barra) para denotar a média de uma amostra.

Independente de se estar trabalhando com uma população ou uma amostra, a média de um conjunto qualquer de dados é definida como sendo a soma de todos os valores observados, dividida pelo número total de observações.

Notação:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n},$$

onde

x_i é o i-ésimo valor observado da variável em estudo;

n é o número total de observações da amostra;

N é o número total de observações da população.

Obs: Raramente se calcula μ uma vez que, na maioria das vezes apenas os dados da amostra são conhecidos. Desde modo, desejando conhecer o valor de μ, calcula-se o valor de $\overline{X}$ e o usa como uma aproximação, ou *estimativa*, de μ.

Exemplo 1:

Sejam os pesos ao nascer, em Kg, de 10 cordeiros da raça Corriedale:

3,2 3,2 2,8 2,1 2,9 3,1 3,2 3,0 3,5 4,0

O peso médio é de 3,1 Kg.

Importante: Nem sempre a média é o valor da variável que ocorre com maior freqüência e, não é, necessariamente, o ponto central da distribuição (ponto que divide as observações exatamente na metade). A média pode ser tomada como o "centro de gravidade", isto é, o ponto de qualquer distribuição em torno do qual se equilibram as discrepâncias (resíduos, afastamento ou desvios) positivas e negativas.

- ***Vantagens e desvantagens da média***

1. É uma medida de tendência central que, por uniformizar os valores de um conjunto de dados, não representa bem os conjuntos que revelam tendências extremas.

2. Não necessariamente tem existência real, isto é, nem sempre é um valor que faça parte do conjunto de dados, para bem representá-lo, embora pertença obrigatoriamente ao intervalo entre o maior e o menor valor.
3. É facilmente calculada.
4. Serve para compararmos conjuntos semelhantes.

1.2. *Mediana* (Md ou $\tilde{X}$)

A mediana de um conjunto de dados ordenados, é o termo do conjunto que o divide em duas partes iguais, isto é, divide o conjunto em dois subconjuntos com o mesmo número de elementos tais que a cada um deles pertencem todos os elementos menores ou todos os elementos maiores que a mediana.

- *n* é ímpar.

Med = valor da variável que ocupa a posição $\frac{n+1}{2}$.

- *n* é par.

Med = média entre os valores da variável que ocupam as posições $\frac{n}{2}$ e $\frac{n+2}{2}$.

Exemplo 2:

Sejam os pesos ao nascer, em Kg, de 10 cordeiros da raça Corriedale:

2,1 2,8 2,9 3,0 3,1 3,2 3,2 3,2 3,5 4,0

O peso mediano é de 3,15 Kg.

Nota: Não é influenciada por valores extremos (é uma medida robusta)

❖ ***Vantagens e Desvantagens da Mediana***

1) Não depende de todos os valores do conjunto de dados, podendo mesmo não se alterar com a modificação.
   Observe, por exemplo, que os conjuntos C e D abaixo possuem o mesmo valor mediano (*Md* = 16), embora sejam bem diferentes.
   Conjunto C : 10, 13, 15, 16, 18, 20 e 21.
   Conjunto D : 1, 10, 13, 16, 18, 20 e 68.
2) Não é influenciada por valores extremos (grandes) do conjunto de dados.
   O valor mediano (38) dos conjuntos E e F abaixo não foi influenciado pelos valores grandes (95 e 120) do conjunto F, o que não acontece com a média.
   Conjunto E : 29, 31, 33, 34, 38, 42, 45, 50 e 51.
   Conjunto F : 29, 31, 33, 34, 38, 42, 45, 95 e 120.
3) Quando há valores repetidos, a interpretação do valor mediano não é tão simples.
   Admitindo como resultado da aplicação de um teste a um conjunto de alunos, as seguintes notas:

   2, 2, 5, 5, 5, 5, 7, 7, 5, 8, 8, 5,

   o valor mediano seria a nota 5 e, no entanto, só existem 2 notas menores e 4 maiores do que 5. Esta desvantagem, unida ao fato da inadequacidade da sua expressão para o manejo matemático, faz com que, em análises estatísticas, a mediana seja menos utilizada do que a média.

1.3. *Moda* (*Mo* ou $\hat{X}$)

Moda de um conjunto de observações é definida como sendo o valor que ocorre com maior freqüência.

De acordo com o comportamento das observações, pode-se ter:

- *Conjunto amodal:* não existe moda, pois todos os valores do conjunto ocorrem com a mesma freqüência. Por exemplo, no conjunto 2, 2, 3, 3, 4, 4, 7, 7, 5 e 5, todos os elementos têm a mesma freqüência (2).
- *Conjunto modal (ou unimodal):* existe uma única moda. Por exemplo, a moda do conjunto 3, 4, 4, 8, 4, 5, 4, 3, 5, 4, 9, 4, 3 e 6, é *Mo* = 4.
- *Conjunto bimodal:* existem duas modas. Por exemplo, o conjunto 3, 5, 5, 5, 5, 10, 10, 10, 10 e 15, é bimodal, pois possui duas modas, *Mo* = 5 e *Mo* = 10.
- *Conjunto multimodal:* existem mais de duas modas. Por exemplo, o conjunto 2, 2, 2, 4, 4, 5, 5, 5, 6, 6, 8, 8, 8, 9 e 10, é multimodal, pois possui três modas, *Mo* = 2, *Mo* = 5 e *Mo* = 8..

Exemplo 3:

Sejam os pesos ao nascer, em Kg, de 10 cordeiros da raça Corriedale:

2,1 2,8 2,9 3,0 3,1 3,2 3,2 3,2 3,5 4,0

A moda é o peso de 3,2 Kg. (unimodal)

❖ ***Vantagens e Desvantagens da Moda***

1) Não depende de todos os valores da série, nem de sua ordenação, podendo mesmo não se alterar com a modificação de alguns deles. Como exemplo, observe as séries A e B.

Série A: 6, 1, 7, 7, 7, 15, 7, 5, 8, 12, 7, 11.

Série B: 11, 7, 7, 12, 5, 5, 56, 7, 7, 58, 60, 7, 15.

As séries A e B possuem a mesma moda (*Mo* = 7), embora sejam bem diferentes.

2) Não é influenciada por valores extremos (grandes) da série. Observe, por exemplo, que a moda (*Mo* = 12) da série abaixo não foi influenciada pelos valores grandes (88, 89 e 100) da série:

Série: 7, 8, 8, 8, 9, 9, 12, 12, 12, 12, 12, 12, 13, 15, 16, 88, 89 e 100.

3) Sempre tem existência real, ou seja, sempre é representada por um elemento do conjunto de dados, excetuando o caso de classes de freqüências, quando trabalhamos com subconjuntos (dados agrupados) e não com cada elemento isoladamente. Veja, por exemplo, que a moda da série abaixo é *Mo* = 15, e 15 é um elemento da série.

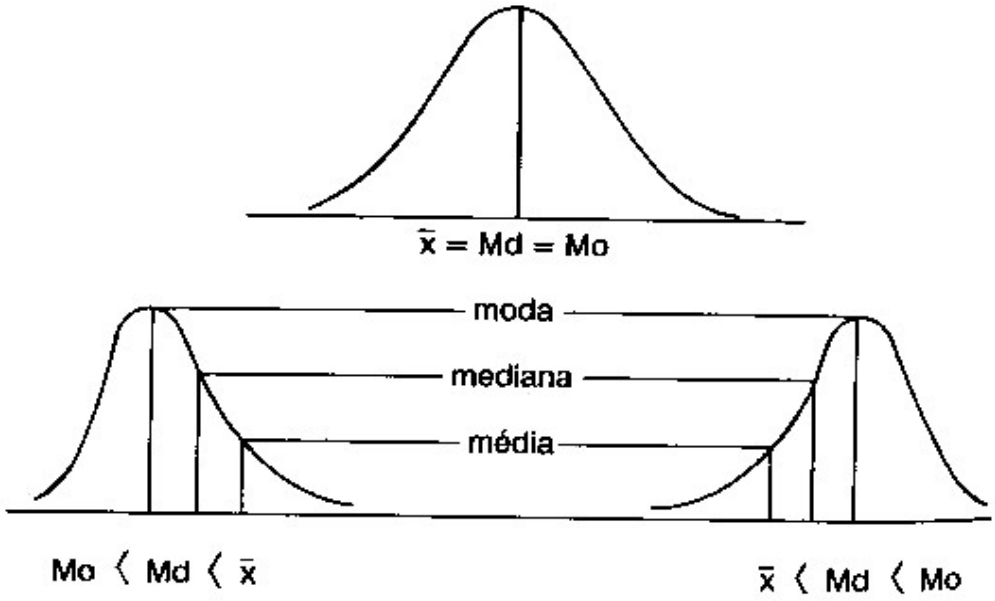
Série: 7, 9, 13, 13, 15, 15, 15, 15, 18 e 20.

2. A forma da distribuição de freqüências e as medidas de tendência central

A distribuição de freqüência da população é um conceito muito importante. Na realidade, raramente se conhece a forma exata da distribuição da população que se deseja estudar. Em geral, se tem apenas uma amostra da população e a partir do histograma (ou polígono de freqüência) dessa amostra é que se obtém alguma idéia sobre a forma da distribuição de freqüência da população.

As medidas de assimetria procuram caracterizar como e quanto a distribuição de freqüências se afasta da condição de simetria. Veja a Figura 1.

Figura 1 - Distribuição de freqüências simétrica, assimétrica positiva e assimétrica negativa.



Importante: Quando realizamos um estudo descritivo, é muito improvável que a distribuição de freqüências seja totalmente simétrica. Na prática, diremos que a distribuição de freqüências é simétrica, caso o seja de um modo aproximado. Por outro lado, ainda observando cuidadosamente o gráfico, podemos não ver claramente de que lado estão as freqüências mais altas. Define-se, então, toda uma família de estatísticas que ajudam a interpretar a assimetria, denominadas **índices de assimetria**.

2.1. Momento central de terceira ordem

Denomina-se momento central de terceira ordem a quantidade:

$$m_3 = \frac{1}{n}\sum (x_i - \bar{X})^3$$

Se $m_3 > 0$, a distribuição é assimétrica positiva (à direita).

Se $m_3 < 0$, a distribuição é assimétrica negativa (à esquerda).

Se $m_3 = 0$, a distribuição é simétrica.

2.2. Índice de assimetria de Pearson

$$I_A = \frac{(\bar{X} - Mo)}{S} \quad \text{ou} \quad I_A = 3\frac{(\bar{X} - Med)}{S} \ ,$$

Deste modo, s

Se $I_A = 0$ a distribuição é simétrica.

Se $I_A > 0$ a distribuição é assimétrica positiva (à direita).

Se $I_A < 0$ a distribuição é assimétrica negativa(à esquerda).

3. Outras Medidas de Posição

3.1. Percentis, Quartis e Decis

Dados que produzem histogramas simétricos são adequadamente descritos e sintetizados pela média e o desvio-padrão. Isso não ocorre em distribuições assimétricas. Quando dizemos que certo aluno está entre os 5% melhores do colégio ou que um país está entre os 10% mais pobres, não precisamos nem saber quantos alunos tem o colégio ou em quantos países estão sendo consideradas as rendas. Aqui já houve uma padronização da posição usando-se a *porcentagem* de alunos ou países com desempenho ou renda *abaixo* do valor considerado. É este raciocínio que define os percentis.

Definição:

O percentil de ordem k (onde k é qualquer valor entre 0 e 100), denotado por Pk, é o valor tal que k% dos valores do conjunto de dados são menores ou iguais a ele. Assim, o percentil de ordem 10, o P10, é o valor da variável tal que 10% dos valores são menores ou iguais a ele; o percentil de ordem 65 deixa 65% dos dados menores ou iguais a ele, etc.

Os percentis de ordem 10, 20, 30, ..., 90 dividem o conjunto de dados em dez partes com mesmo número de observações e são chamados de *decis*.

Os percentis de ordem 25, 50 e 75 dividem o conjunto de dados em quatro partes com o mesmo número de observações.

Existem vários processos para calcular os percentis, usando interpolação. Vamos utilizar um método mais simples. As diferenças serão muito pequenas e desaparecerão à medida que aumenta o número de dados.

De modo geral, para se obter o percentil de ordem k, denotado por Pk, após ordenar os dados, calcula-se o valor $L = (\frac{k}{100})n$. Se L for inteiro, o valor do Pk é a média entre o L-ésimo e o (L+1)-ésimo valores a contar do menor. Se L não for inteiro, arredonde L para o maior inteiro mais próximo, e o valor de Pk será o L-ésimo valor a contar do menor.

Exemplo 4: Considere os pesos, em Kg, de 40 borregas e ovelhas da raça Hampshire Down, já colocados em ordem crescente:

40 41 42 42 44 47 48 48 49 49 51 52 53 58 59 62 63 64 65 66
67 68 69 70 75 76 83 83 85 86 86 87 87 88 92 93 94 95 97 98

Percentil de ordem 10: 10% de 40 = 4. Então o P10 = média(4o e 5o valores)=(42+44)/2 = 43.
Percentil de ordem 95: 95% de 40 = 38. Então o P95 = média(38o e 39o valores)=(95+97)/2 = 96.
Primeiro Quartil: 25% de 40 = 10. Então o Q1 = média(10o e 11o valores)=(49+51)/2 = 50.
Terceiro Quartil: 75% de 40 = 30. Então o Q3 = média(30o e 31o valores)=(86+86)/2 = 86.
Mediana: 50% de 40 = 20. Então mediana = média(20o e 21o valores)=(66+67)/2 = 66,5.

4. Estatísticas de achatamento

Podemos ter interesse em saber se a distribuição dos dados é mais ou menos achatada (comprida e estreita). Esse achatamento é medido em comparação com uma certa distribuição de freqüências que consideraremos NORMAL (não por casualidade, é esse o nome que recebe a distribuição de referência).

4.1. Coeficiente de achatamento de Fisher (Curtose)

Define-se o coeficiente de achatamento de Fisher (Curtose) como:

$$\gamma_2 = \frac{m_4}{s^4} - 3 \ ,$$

em que, m_4 é o momento central de quarta ordem dado por $m_4 = \frac{1}{n}\sum (x_i - \bar{x})^4$.

- é um coeficiente adimensional, invariante perante trocas de escala e de origem;
- serve para medir se uma distribuição de frequências é muito achatada ou não.

Para afirmar que a distribuição é comprida ou estreita, deve-se ter um padrão de referência, que é a distribuição NORMAL ou GAUSSIANA, para a qual se tem:

$$\frac{m_4}{s^4} = 3 \Rightarrow \gamma_2 = 0$$

Desse modo, de acordo com γ_2 classificam-se as distribuição de frequências em:

Leptocúrtica: quando $\gamma_2 > 0$, ou seja, quando a distribuição de frequências é mais achatada que o normal;

Mesocúrtica: quando , $\gamma_2 = 0$ ou seja, quando a distribuição de frequências é tão achatada quanto o normal;

Platicúrtica: quando $\gamma_2 < 0$, ou seja, quando a distribuição de frequências é menos achatada que o normal.

5. Resumo dos 5-Números

O resumo de 5-números associa os limites inferior e superior do conjunto de dados aos quartis, fornecendo uma idéia bastante razoável da dispersão, da tendência central e da forma da distribuição, isto é, do grau de deformação.

5.1. Box-Plot : É uma representação gráfica dos dados através de seu resumo de 5-números.

O Boxplot fornece informações importantes sobre o comportamento dos dados, como a simetria e variabilidade, e auxilia na detecção de *outliers*.

Para sua construção é necessário ter:

O primeiro quartil (Q_1)

A mediana (Med)

O terceiro quartil (Q_3)

O desvio interquartílico ($DQ = Q_3 - Q_1$)