UFES
Universidade Federal do Espírito Santo - UFES
Centro de ciências agrárias e engenharias - CCAE
Departamento de Engenharia Rural

ESTATÍSTICA DESCRITIVA

- CADERNO DIDÁTICO -

GUSTAVO SESSA FIALHO[1]
MAICON NARDINO[2]
EDVALDO FIALHO DOS REIS[3]
JOSÉ FRANCISCO TEIXEIRA DO AMARAL[3]

[1]Departamento de Matemática e Estatística, Universidade Federal de Pelotas (UFPEL), Campus Capão do Leão, Prédio 05, Sala 305A, 96010-900, Pelotas, Rio Grande do Sul, Brasil.
[2]Departamento de Fitotecnia, Universidade Federal de Viçosa (UFV), Viçosa, Minas Gerais, Brasil.
[3]Departamento de Engenharia Rural, Universidade Federal do Espírito Santo (UFES), Alegre, Espírito Santo, Brasil.

Tratamento da Informação - Estatística

INTRODUÇÃO

A estatística é um ramo da ciência de ampla aplicação em quase todas as áreas do conhecimento. Aprender estatística significa bem mais que usar *softwares* sofisticados. Antes porém, faz-se necessário a aquisição de embasamento teórico que permita a coleta, tratamento e consequente interpretação da informação contida em um conjunto de dados. Assim, o universo em estudo poderá ser melhor compreendido permitindo a tomada de decisões razoáveis no processo organizacional.

A estatística pode ser dividida em duas grandes áreas: estatística descritiva e estatística indutiva ou inferencial.

A descritiva destina-se a descrever, sintetizar e organizar um conjunto de dados, principalmente, quanto a medidas de tendência central e variabilidade; permitindo assim, que se tenha ampla compreensão da representatividade dos mesmos. Por outro lado, nenhuma generalização ou inferência sobre a população de origem os destes dados é realizada.

A indutiva ou inferencial propõe-se a fazer generalizações sobre uma população à partir da análise de uma "amostra" representativa seus de dados. As premissas para a validação destas análises é parte fundamental no escopo teórico da estatística indutiva.

Este Texto tem como objetivo tratar da estatística descritiva, em particular, da área que versa sobre tratamento da informação.

ESTATÍSTICA DESCRITIVA

A Estatística Descritiva tem como objetivo coletar, organizar, resumir, analisar, interpretar e apresentar um conjunto de dados de forma adequada (Figura 1). Para tal, torna-se essencial uma análise de qualidade que possibilite a detecção de tendências a eles associadas. Isto permite a padronização e comparações com outros resultados, além de dar base para a validação do modelo a ser usado no tratamento desses dados.

A Estatística Descritiva pressupõe o seguinte fluxograma:

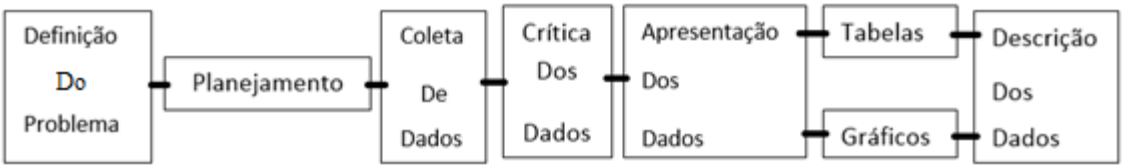

Figura 1: Fluxograma da descrição estatística de um conjunto de dados coletados mediante ao estudo de um problema proposto

Fonte: Regazzi (1997).

Após o planejamento da pesquisa, procede-se a coleta de dados. Posteriormente, os mesmos devem ser criticados objetivando-se eliminar erros capazes de provocar equívocos, valores estranhos ao levantamento devem estudados e sendo o caso, suprimidos. Após a crítica convém apresentá-los racionalmente, o que facilita o entendimento do fenômeno em estudo e pode ser feito através de tabelas e gráficos. Inclui-se também nesta parte, a descrição dos dados por meio de medidas que os sumarizem. Elas podem ser: medidas de posição, separatrizes, dispersão, assimetria e curtose.

DEFINIÇÕES E CONCEITOS FUNDAMENTAIS

Em estatística, dados podem ser coletados diretamente na população ou em uma amostra dessa população. Assim, define-se:

- população é um conjunto de elementos que apresentam pelo menos uma característica em comum. A população é considerada o conjunto universo ou conjunto de interesse a ser trabalhado; A caraterização da população se dá através de constantes a ela inerentes, parâmetros populacionais.

- amostra é um subconjunto do conjunto população, isto é, todo elemento da amostra também é um elemento da população, entretanto, são conjuntos distintos. A amostra deve ser construída de tal forma que seja representativa da população.

Nota: Como se tratam de conjuntos, estes podem ser finitos ou infinitos; e

- variável é a quantidade a ser observada, medida ou contada na população, ou amostra, e que apresenta variação de elemento para elemento.

As variáveis podem ser assim classificadas:

- **variável quantitativa** é uma variável cujos valores representam quantidades. Estas quantidades podem ser contínuas, isto é, podem assumir qualquer valor numérico na reta numérica real. Como exemplo, a idade dos alunos que cursam o CLMD (tempo é uma variável contínua e pode ser medida com a precisão que se deseja). Naturalmente, as variáveis contínuas constituem um conjunto infinito e não enumerável. Também uma variável quantitativa pode ser discreta, isto é, só pode assumir valores inteiros, formando um conjunto finito ou infinito enumerável. Como exemplo, o número de alunos matriculados no CLMD;

-variável qualitativa ou categórica é uma variável cujos valores não são quantidades, mas sim qualidades ou atributos. Estas variáveis são usualmente codificadas usando-se números, entretanto tais números somente significam a representação da variável categórica, sem conotação quantitativa. Se há ordenação, isto é, existe uma sequência natural de ordem entre as categorias, a variável qualitativa ou categórica é dita ordinal. Por exemplo, o desempenho acadêmico dos alunos matriculados no CLMD (escala semântica: insuficiente, regular, bom, ótimo). Por outro lado, se não há uma ordenação natural das categorias, a variável categórica é dita nominal. Como exemplo, o sexo (masculino ou feminino).

Nota: É frequente, para melhor organização e apresentação dos dados, a categorização das variáveis quantitativas. Como exemplo, a variável quantitativa idade pode ser categorizada em faixas etárias (0-10, 11-20, 21-30, ...), caracterizando, assim, a variável como qualitativa ordinal em que há uma sequência de ordem entre as categorias representadas pelas faixas. Também uma variável categórica, pode ser representada por um número, por exemplo, o sexo pode ser codificado como masculino (1) e feminino (2).

-parâmetros são constantes populacionais, por exemplo entre outros, a média populacional.

-estimadores são estatísticas utilizadas em uma amostra para o cálculo de estimativas dos parâmetros da população que a originou.

Tabela 1: Principais parâmetros e seus respectivos estimadores

Estimadores pontuais dos principais parâmetros populacionais, entre outros

	PARÂMETRO	ESTIMADOR
Média	μ	$\bar{X}$
Variância	σ^2	$\hat{\sigma}^2$
Desvio Padrão	σ	$\hat{\sigma}$

ANÁLISE EXPLORATÓRIA DE DADOS

Medidas de tendência Central ou de Medidas de Posição

São estatísticas representativas da localização ou do posicionamento dos valores de uma amostra de dados ao longo da escala de medidas. As mais conhecidas e utilizadas são a média, a mediana e a moda, que numa distribuição simétrica são coincidentes. Nesta categoria também se enquadram os quartis, decis e percentis.

Para a construção das definições a seguir, consideremos uma variável X assumindo particulares valores: $X = \{x_1, x_2, x_3, ..., x_n\}$, em que n representa o tamanho da amostra de dados da referida variável.

Média Aritmética

A medida de tendência central mais utilizada é a média aritmética, denotada por $\bar{X}$. Ela consiste na soma de todas as observações dos valores amostrais dividida pelo tamanho da amostra; sendo dada por:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Caso os citados valores estejam associados às suas respectivas frequências: f_1, f_2, f_3 ..., f_n; a média aritmética será:

$$\bar{X} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_n x_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i}^{n} f_i}.$$

Propriedades da média aritmética

P_1: somando-se ou subtraindo-se uma constante a cada um dos valores assumidos pela variável, a média aritmética fica somada ou subtraída dessa constante.

Prova: seja $X = \{x_1+a, x_2+a, x_3+a, ..., x_n+a\}$, em que a é uma constante. A média aritmética de X será:

$$\bar{X} = \frac{(x_1 + a) + (x_2 + a) + (x_3 + a) + \cdots + (x_n + a)}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} a}{n} = \frac{\sum_{i=1}^{n} x_i}{n} + \frac{na}{n} = \bar{X} + a$$

P_2: multiplicando-se ou dividindo-se uma constante a cada um dos valores assumidos pela variável, a média aritmética fica multiplicada ou dividida pela constante.

Prova: seja $X = \{ ax_1, ax_2, ax_3, ..., ax_n\}$, em que a é uma constante. A média aritmética de X será:

$$\bar{X} = \frac{ax_1 + ax_2 + ax_3 + \cdots + ax_n}{n} = \frac{a(x_1 + x_2 + x_3 + \cdots + x_n)}{n} = a\frac{\sum_{i=1}^{n} x_i}{n} = a\bar{X}$$

P_3: a soma algébrica desvios dos valores assumidos pela variável em relação à média aritmética é nula.

Prova: seja o i-ésimo desvio do i-ésimo valor assumido por X em relação a $\bar{X}$, $d_i = x_i - \bar{X}$. A soma algébrica dos possíveis desvios X será:

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (x_i - \bar{X})$$

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{X} = \sum_{i=1}^{n} x_i - n\bar{X} = \sum_{i=1}^{n} x_i - n\frac{\sum_{i=1}^{n} x_i}{n} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

P_4: a soma dos quadrados dos desvios em relação à média aritmética é um valor mínimo.

Prova:

$$SQD_X = \sum_{i=1}^{n} (x_i - \bar{X})^2 = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\bar{X}x_i + \sum_{i=1}^{n} \bar{X}^2$$

Derivando SQD_X em relação a $\bar{X}$ e igualando a derivada a zero, encontra-se o ponto ótimo, assim:

$$\frac{\partial SQD_X}{\partial \bar{X}} = -2\sum_{i=1}^{n} x_i + 2n\bar{X} = 0$$

$$2n\bar{X} = 2\sum_{i=1}^{n} x_i$$

Logo, se a segunda derivada da SQD_X em relação a $\bar{X}$ for positiva o valor da SQD_X é um mínimo. Desta forma,

$$\frac{\partial SQD_X}{\partial \bar{X}\partial \bar{X}} = 2n > 0$$

que para qualquer tamanho de amostra será positivo.

Torna-se importante salientar que todos os valores observados da variável X participam do cálculo da média aritmética, o que não ocorre com todas as medidas de tendência central. Destaca-se também que valores amostrados muitos discrepantes tendem a exercer influência desproporcional sobre a média e que, para qualquer conjunto de dados será sempre possível calcular a média aritmética, sendo a mesma, estimativa única para o referido conjunto.

A média aritmética é de fácil interpretação e também é o ponto de equilíbrio de uma distribuição de um conjunto de observações. Isto concorre para que a mesma seja uma medida descritiva tão eficiente quanto mais simétrica for a distribuição das observações ao seu redor.

Mediana (Md)

A mediana é o valor central de uma amostra de elementos dispostos em rol, aquele que divide a amostra em duas subamostras de mesmo tamanho, neste caso *n* é ímpar. Se *n* for par, a mediana será a média aritmética dos dois valores centrais.

A mediana é preferível à média quando existem valores discrepantes na amostra de dados. Se a amostra é simétrica, então a mediana será igual ou próxima ao valor da média. Assim, considerando os dados em rol:

*amostra de tamanho ímpar: (Md) = elemento que ocupa a posição $\frac{n+1}{2}$;

*amostra de tamanho par: (Md) é a média aritmética dos valores que ocupam as posições $x_{\frac{n}{2}}$ e $x_{\frac{n+2}{2}}$.

Quantis

A amostra de dados ordenada pode também ser dividida por meio das seguintes medidas: quartil, decil e percil. Os quartis Q1, Q2 e Q3, dividem a amostra de dados em quatro partes iguais, tendo cada parte 25% dos dados. Os decis D1, D2, ... D5, ..., D9 dividem a amostra em dez partes iguais, tendo cada uma 10% dos dados e os

percentis P1, P2, ..., P50 ..., D99 dividem a amostra em cem partes iguais, cada um com 1% (Tabela 2).

Tabela 2: Equivalência entre os principais quantis utilizados na descrição de dados

(%) dos dados abaixo de um quantil	Quartil	Decil	Percentil	Mediana
25	Q_1	$D_{2,5}$	P_{25}	
50	Q_2	D_5	P_{50}	Md
75	Q_3	$D_{7,5}$	P_{75}	

Existem diferentes estatísticas para a determinação dos quantis. Apresentar-se-á uma que se baseia na média ponderada. Assim:

$$Q_{(p)} = (1 - g) * x_{(j)} + g * x_{(j+1)}$$

em que:

p é o t-ésimo quantil fixado entre 0 e 1 dado por:

- para os quartis: $p = t/4$;
- para os decis: $p = t/10$;
- para os percentis: $p = t/100$;

j é a parte inteira de $(n + 1)p$;

g é a parte decimal de $(n + 1)p$; e

n é o tamanho da amostra.

Moda

A moda é o valor que ocorre com maior frequência em um conjunto de dados, ou seja, o valor que mais se repete. Não havendo observações de mesma frequência a amostra será classificada como amodal. Por outro lado, será bimodal, trimodal ou multimodal se apresentar duas, três ou mais de três modas, respectivamente.

Geralmente a moda é utilizada quando se quer conferir destaque a um valor de alta frequência em um conjunto de observações. Salienta-se ainda que, para amostras cujas observações sigam a distribuição normal, a média, a mediana e a moda assumem os mesmos valores (Figura 2 – B).

Essa coincidência deixa de existir em distribuições assimétricas. Desta forma: Moda ≤ Mediana ≤ Média.

Medidas de dispersão

São medidas utilizadas para quantificar o grau de variedade dos valores de uma amostra de dados em torno da sua média.

Amplitude Total

Trata-se da diferença entre o maior e o menor valor observado na amostra de dados. Prestando-se muito bem para uma avaliação preliminar do conjunto amostrado. Se a variável em estudo apresenta estremos conhecidos, particulares valores discrepantes na variável podem, assim, ser identificados. A amplitude total também indica que o desvio entre duas observações quaisquer será no máximo igual a AT.

$$AT_X = x_{máx} - x_{min}$$

Variância

A variância ($\hat{\sigma}^2$) quantifica a dispersão dos valores das observações de um conjunto de dados em torno de sua média ($\bar{X}$). Obtém-se ($\hat{\sigma}_X^2$) pela soma dos quadrados dos desvios de cada valor de (x_1, x_2, x_3, ..., x_n) em relação a ($\bar{X}$), dividida pelo número de graus de liberdade da amostra. Desta forma, ($\hat{\sigma}_X^2$) é a média dos (n-1) desvios quadráticos e independentes. A variância é um termo ao quadrado, Assim:

$$\hat{V}(X) = \hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1}$$

A variância comporta ainda uma expressão mais prática, expressa em termos da esperança matemática.

$$V(X) = E[X - E(X)]^2$$
$$V(X) = E(X^2) - [E(X)]^2$$

- Propriedades da variância:

a) a variância de uma constante é zero:

$$V(k) = E[k - E(k)]^2 = E[k - k]^2 = O$$

b) somando-se ou subtraindo-se uma constante K a uma variável aleatória (v.a), a variância não se altera:

$$\begin{aligned} V(X \pm k) = \ & E[(X \pm k) - E(X \pm k)]^2 \\ & = E[(X) - E(X) \pm (k - k) - E(X)]^2 \\ & = E[X - E(X)]^2 = V(X) \end{aligned}$$

c) multiplicando-se uma v.a por uma constante k sua variância fica multiplicada pelo quadrado da constante:

$$V(kX) = k^2V(X)$$

d) a variância da soma ou subtração de duas variáveis independentes é igual a soma das variâncias das variáveis, isso porque a covariância entre elas será nula:

$$V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$$

$$V(X - Y) = V(X) + V(Y) - 2Cov(X, Y)$$

Desvio de padrão

O desvio padrão de uma variável X ($\hat{\sigma}_X$) é a raiz quadrada positiva da variância ($\hat{\sigma}_X^2$). Tal medida é utilizada para retornar a à unidade original da variável a medida de dispersão permitindo com que a mesma adquira melhor interpretação. Assim:

$$\hat{\sigma}_X = \sqrt{\hat{\sigma}_X^2}$$

Nota-se que quanto mais homogênea for a amostra, valores próximos uns dos outros, menor será o desvio padrão.

Coeficiente de variação (CV)

É o desvio padrão amostral expresso em porcentagem da média. Assim:

$$CV_{\%} = \frac{\hat{\sigma}_X}{\bar{X}} * 100$$

O CV, por ser uma medida adimensional, presta-se muito bem para comparações referentes à variabilidade de amostras distintas que apresentem médias muito desiguais e/ou unidades de medida diferentes.

Erro padrão da média

É uma medida que expressa a precisão ou a representatividade da estimativa obtida para a média. Assim:

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}_X}{\sqrt{n}}$$

Em um levantamento, é comum apresentar a média acompanhada de seu erro-padrão, $\bar{X}_{\pm\hat{\sigma}_{\bar{X}}}$.

Medidas de forma: assimétrica e curtose

Assimetria

É uma medida que reflete o grau de afastamento ou desvio da simetria da distribuição de um conjunto de dados.

A distribuição será assimétrica positiva, quando for inclinada para direita e apresentar maior número de observações amostrais menores que a média (Figura 2 – A), neste caso:

$$Mo < Md < \bar{X}$$

De modo oposto, se a distribuição inclina-se para a esquerda, assimétrica negativa, haverá mais valores amostrais maiores que a média e a calda da distribuição se alongará no sentido dos valores menores que a média, (Figura 2 – C), neste caso:

$$\bar{X} < Md < Mo$$

Em uma distribuição simétrica (Figura 2 - B), ter-se-á:

$$\bar{X} = Md = Mo$$

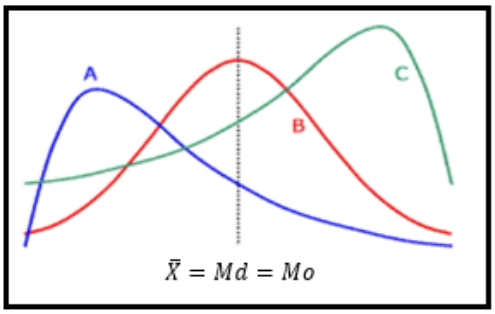

Figura 2: Assimetria das curvas de distribuição: assimétrica positiva (A), simétrica (B) e assimétrica negativa (C).

A estimativa do coeficiente de assimetria ($\hat{s}$) de uma variável X é dada por:

$$\hat{s} = \frac{n}{(n-1)*(n-2)} * \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{\hat{\sigma}_x} \right)^3$$

Se o resultado for zero a distribuição será simétrica se for negativo ou positivo será assimétrica negativa ou positiva, respectivamente.

Curtose

É uma medida que reflete o grau de achatamento da distribuição os dados. Para se estimar o grau de curtose ($\hat{k}$) de uma variável (X) utiliza-se o seguinte estimador:

$$\hat{k} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} * \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{\hat{\sigma}_x} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Se o resultado for zero a distribuição será a normal sendo chamada de mesocúrtica; se negativo, a distribuição será achatada indicando alta variabilidade no conjunto de dados e chamada de planicúrtica e; se positivo a distribuição será centrada

em torno da média indicando alta homogeneidade e chamada de leptocúrtica (Figura 3).

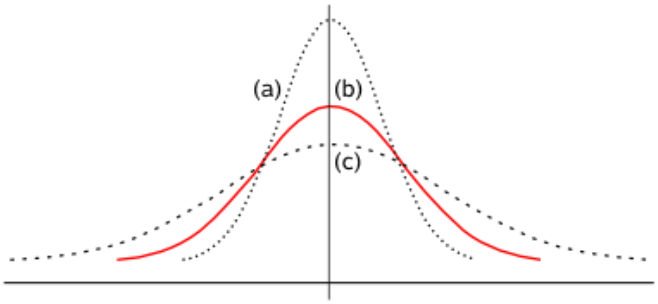

Figura 3: Curva leptocúrtica (a), mesocúrtica (b) e planicúrtica (c).

DISTRIBUIÇÃO DE FREQUÊNCIAS

No estudo de um grande conjunto de dados é importante organizá-los e resumi-los em uma tabela, agrupando-os em classes com suas respectivas frequências.

Se as observações referem-se a dados discretos, sua identificação e contagem representam as classes. Por outro lado, se forem contínuos, as classes serão estabelecidas por um intervalo de classe (h).

Intervalo de Classe - h

É calculado dividindo-se a amplitude total (AT), pelo número de classes (k). Assim:

$$h = \frac{AT}{k}$$

Assim h representará o valor encontrado da diferença entre o limite superior (LS) e inferior de cada classe (LI). A representação dos limites de classe é bastante variável,

entretanto uma das mais utilizadas é a que considera o intervalo fechado à esquerda e aberto à direita, desta forma, a classe incluirá as frequências dos valores compreendidos entre o LI e o LS, incluindo *as* do LI e excluindo as do LS.

Amplitude Total – AT

É a diferença entre os valores da maior e da menor observação encontrados no conjunto de dados sob análise. Assim:

$$AT = Obs_{max} - Obs_{min}$$

Número de Classes - k

A definição do números de classes afeta diretamente a correta interpretação do resumo do conjunto de dados sob análise. Vê-se que, quanto maior for o número de classes, mais pulverizada torna-se a informação, ao passo que, quanto menor, a variabilidade inerente ao conjunto resumir-se-á em poucas classes, tornando a distribuição de frequências pouco informativa.

Neste escopo, torna-se fundamental a experiência e o bom senso do pesquisador referente ao tamanho da amostra de trabalho e ao número ideal de classes para a sua representação em uma tabela de distribuição de frequências; assim, poder-se-á observar de fato como os valores se distribuem.

Uma das regras práticas mais utilizadas na definição do número de classes é a regra de *Sturges*. Em que:

$$k = 1 + 3{,}22 * log_n$$

Passos para a construção de uma tabela de distribuição de Frequências

1 – organizar o conjunto de dados brutos em rol;

2 – estabelecer o número de classes bem como o intervalo de agrupamento;

3 – enquadrar as observações do conjunto nas respectivas classes mediante contagem; e

4 – finalizar com a construção da tabela de frequências.

Tabela genérica de distribuição de frequências

Abaixo segue a tabela genérica de distribuição de frequências segundo Ribeiro-Júnior (2008):

Tabela 3 – Frequências de k classes obtidas em n observações

	Classes	f_i	f_{ai}	f_{ri}	f_{ari}	PM_i
1	$LI_1 \vdash LS_1$	f_1	f_1	f_1/F_t	f_{a1}/F_t	(LI1+LS1)/2
2	$LI_2 \vdash LS_2$	f_2	$f_1 + f_2$	f_2/F_t	f_{a2}/F_t	(LI2+LS2)/2
3	$LI_3 \vdash LS_3$	f_3	$f_1 + f_2 + f_3$	f_3/F_t	f_{a3}/F_t	(LI3+LS3)/2
...	...	...	...	...	...	...
k	$LI_k \vdash LS_k$	f_k	F_t	f_k/F_t	1,0	(LIk+LSk)/2

f_i – frequência simples da classe i;

f_{ai} – frequência acumulada da classe i;

f_{ri} – frequência relativa da classe i;

f_{ari} – frequência acumulada relativa da classe i; e

PM_i – ponto médio da classe i

Medidas de posição para dados agrupados em tabelas de distribuição de frequência

Média aritmética

Quando os dados estiverem agrupados em tabelas de distribuição de frequência, a média é estimada ponderando-se o valor médio da classe por sua respectiva frequência. Deste modo,

- para dados discretos:

$$\bar{X} = \frac{\sum f_i x_i}{n}$$

- para dados contínuos

$$\bar{X} = \frac{\sum f_i \bar{X}_i}{n}$$

onde: f_i e $\bar{X}_i$ representam respectivamente, a frequência simples e o ponto médio da classe i.

Mediana

Para que se possa estimar a mediana a partir de dados agrupados é necessário definir a classe mediana e em seguida executar uma interpolação. Neste caso, a classe mediana será aquela que contiver frequência acumulada igual ou imediatamente superior a n/2. Assim,

$$Md = LI_{Md} + \frac{(0{,}5n - f_{acaMd})}{f_{iMd}} * h_{Md}$$

onde LI_{Md}, f_{iMd} e h_{Md} referem-se ao limite inferior, frequência simples e amplitude da classe mediana; f_{acaMd} é a frequência acumulada da classe anterior à classe mediana. É importante salientar que se a primeira classe for a classe mediana, $f_{acaMd} = 0$.

Moda

A moda também é obtida por interpolação. Para tanto, torna-se necessário que se defina a classe modal como aquela que contiver a maior frequência simples. Assim,

$$Mo = LI_{Mo} + \left(\frac{f_{iaMo} - f_{iMO}}{f_{iaMo} + f_{ipMo} + 2f_{iMO}}\right) * h_{MO}$$

em que LI_{Mo}, f_{iMO} e h_{MO} referem-se ao limite inferior, frequência simples e amplitude da classe modal; f_{iaMo} e f_{ipMo} são respectivamente a frequência simples anterior e a posterior à da classe modal.

Quantis

Quartil

$$Q_k = LI_{Qk} + \frac{(k * \frac{n}{4} - f_{acaQk})}{f_{iQk}} * h_{Qk}$$

onde LI_{Qk}, f_{iQk} e h_{Qk} referem-se ao limite inferior, frequência simples e amplitude da classe que contém o quartil de ordem k; f_{acaQk} é a frequência acumulada da classe anterior à classe que contém o quartil de ordem k. É importante salientar que se a primeira classe for a classe que contém o referido quantil, $f_{acaQk} = 0$.

Decil

$$D_k = LI_{Dk} + \frac{(k * \frac{n}{10} - f_{acaDk})}{f_{iDk}} * h_{Dk}$$

onde LI_{Dk}, f_{iDk} e h_{Dk} referem-se ao limite inferior, frequência simples e amplitude da classe que contém o decil de ordem k; f_{acaDk} é a frequência acumulada da classe

anterior à classe que contém o decil de ordem k. É importante salientar que se a primeira classe for a que contém o referido quantil, $f_{acaDk} = 0$.

Percentil

$$P_k = LI_{Pk} + \frac{(k * \frac{n}{100} - f_{acaPk})}{f_{iPk}} * h_{Pk}$$

onde LI_{Pk}, f_{iPk} e h_{Pk} referem-se ao limite inferior, frequência simples e amplitude da classe que contém o percentil de ordem k; f_{acaPk} é a frequência acumulada da classe anterior à classe que contém o decil de ordem k. É importante salientar que se a primeira classe for a que contém o referido quantil, $f_{acaDk} = 0$.

Medidas de dispersão para dados agrupados em tabelas de distribuição de frequência

Variância

O estimador para o cálculo da variância para dados agrupados em tabelas de distribuição de frequências será:

$$\hat{V}(X) = \hat{\sigma}_X^2 = \frac{\sum_{i=1}^{n} \hat{d}_i^2 f_i}{\sum_{i=1}^{n} f_i}$$

em que: $\hat{d}_i = \bar{X}_i - \bar{X}$; $\bar{X}_i$ é ponto médio da classe i; $\bar{X}$ média aritmética dos dados agrupados e f_i é a frequência simples da classe i.

Desvio Padrão

Estima-se o desvio padrão extraindo a raiz quadrada da variância demonstrada no tópico anterior.

EXERCÍCIOS DE APLICAÇÃO

1) Para X = {10, 25, 14, 32, 42, 25, 13}, determine:

a) média

b) mediana

c) moda

d) variância

e) desvio padrão

f) coeficiente de variação

g) erro-padrão da média

h) Encontre a mediana utilizando o estimador de quantis apresentado

a)

$$\bar{X} = \frac{(10 + 25 + 14 + 32 + 42 + 25 + 13)}{7} = 23$$

b)

Classificando o conjunto em ordem crescente

10	13	14	**25**	25	32	42
x_1	x_2	x_3	x_4	x_5	x_6	x_7

$$Md = x_4 = 25$$

c)

x_i	10	31	14	**25**	32	42
f_i	1	1	1	**2**	1	1

x_i = Particular valor da variável X

f_i = frequência simples observada

$$Mo = 25$$

d)

$$\hat{\sigma}^2 = \frac{(10^2 + 25^2 + 14^2 + 32^2 + 42^2 + 25^2 + 13^2) - \frac{(161)^2}{7}}{6} = 133{,}3333$$

e)

$$\hat{\sigma} = \sqrt{133{,}3333} = 11{,}5470$$

f)

$$CV_{\%} = \frac{11{,}5470}{23} * 100 = 50{,}20\ \%$$

g)

$$\hat{\sigma}_{\bar{X}} = \frac{11{,}5470}{\sqrt{7}} = 4{,}3644$$

h)

10	13	14	25	25	32	42
x_1	x_2	x_3	x_4	x_5	x_6	x_7

A mediana é igual a Q_2. Assim:

para Q_2, p = 2/4 = 0,5

(n+1)p = (7+1) 0,5 = 4,0

j = 4

g = 0

Q_2= (1-0) * x_4 + 0 * x_5 = x_4 = 25

2) Para Y = {3, 28, 43, 79, 80, 13}, determine:

a) média

b) mediana

c) moda

d) variância

e) desvio padrão

f) coeficiente de variação

g) erro-padrão da média

h) assimetria

i) curtose

j) o percentil 90,5

k) a mediana pelo estimador de quantis

a)

$$\bar{X} = \frac{(3 + 28 + 43 + 70 + 80 + 13)}{6} = 41$$

b)

Classificando o conjunto em ordem crescente

3	13	28	43	79	80
x_1	x_2	x_3	x_4	x_5	x_6

$$Md = \frac{x_3 + x_4}{2} = \frac{28 + 43}{2} = 35{,}5$$

c)

x_i	3	13	28	43	70	80
f_i	1	1	1	1	1	1

$x_i =$ Particular valor da variável X

$f_i =$ frequência simples observada

$$Conjunto\ amodal$$

d)

$$\hat{\sigma}^2 = \frac{(3^2 + 13^2 + 28^2 + 43^2 + 70^2 + 80^2) - \frac{(246)^2}{6}}{5} = 1073{,}2$$

e)

$$\hat{\sigma} = \sqrt{1073{,}2} = 32{,}7597$$

f)

$$CV_{\%} = \frac{32{,}7597}{41} * 100 = 79{,}9018\ \%$$

g)

$$\hat{\sigma}_{\bar{X}} = \frac{32{,}7597}{\sqrt{6}} = 13{,}3741$$

h)

$$\hat{s} = \frac{n}{(n-1)*(n-2)} * \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)^3$$

Para facilitar

x_i	$\left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)$	$\left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)^3$
3	-1,15996	-1,56074
13	-0,85471	-0,62439
28	-0,39683	-0,06249
43	0,061051	0,000228
79	1,159962	1,560742
80	1,190487	1,687229
	$\sum\left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)^3$	**1,000579**

$$\hat{s} = \frac{6}{(6-1)*(6-2)} * 1{,}000579 = 0{,}3002$$

- Distribuição Assimétrica positiva.

i)

$$\hat{k} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} * \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Para facilitar

x_i	$\left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)$	$\left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)^4$
3	-1,15996	1,810401
13	-0,85471	0,53367
28	-0,39683	0,024798
43	0,061051	1,39E-05
79	1,159962	1,810401
80	1,190487	2,008625
	$\sum\left(\frac{x_i - \bar{X}}{\hat{\sigma}_x}\right)^4$	**6,187907**

$$\hat{k} = \left(\frac{6(6+1)}{(6-1)(6-2)(6-3)} * \mathbf{6,187907}\right) - \frac{3(6-1)^2}{(6-2)(6-3)}$$

$$\hat{k} = \left(\frac{42}{60} * \mathbf{6,187907}\right) - \frac{75}{12} = \mathbf{-1,9185}$$

- Distribuição planicúrtica indicando alta variabilidade no conjunto de dados.

j)

3	13	28	43	79	80
x_1	x_2	x_3	x_4	x_5	x_6

para $P_{90,5}$, p = 90,5/100 = 0,905

(n+1)p = (6+1) 0,905 = 6,335

j = 6

g = 0,335

$P_{90,5}$= (1-0,335) * x_6 + 0,335 * x_7

como x_7 não existe, x_7 = 0

$P_{90,5}$= 0,9665 * x_6 = 0,9665 * 80 = 77,32

k) a mediana é igual ao Percentil 50

para P_{50}, p = 50/100 = 0,5

(n+1)p = (6+1) 0,5 = 3,5

j = 3

g = 0,5

P_{50}= (1-0,5) * x_3 + 0,5 * x_4

P_{50}= 0,5 * 28 + 0,5 * 43 = 35,5 = **Md**

3) Qual dos conjuntos acima é o mais homogêneo? X ou Y?

- R: Conjunto X, pois apresenta menor coeficiente de variação.

4) Os dados abaixo referem-se às produções de 25 plantas de uma progênie F2 de soja em g/planta avaliadas no DME/UFPEL 2015 (Obs.: dados fictícios).

3,07	13,73	16,01	19,78	24,4
5,09	14,39	18,27	20,52	28,17
6,14	14,52	18,99	21,51	31,02
8,93	14,72	19,01	23,88	31,2
13,59	15,56	19,78	24,09	35,3

Pede-se:

a) construa uma tabela de distribuição de frequência adotando k=6.

b) calcule a média aritmética para os dados agrupados;

c) calcule a mediana para os dados agrupados;

d) identifique a classe modal para os dados agrupados;

e) calcule a variância amostral para os dados agrupados;

R:

a)

- determinação do intervalo de classe (h)

n	25
x_{max}	35,3
x_{Min}	3,07
AT	32,23
K	6
h	5,37

- Contagem das frequências com base no estabelecimento das classes

3,07	13,73	16,01	19,78	24,4
5,09	14,39	18,27	20,52	28,17
6,14	14,52	18,99	21,51	31,02
8,93	14,72	19,01	23,88	31,2
13,59	15,56	19,78	24,09	35,3

- Construção da Tabela

Tabela 1: Distribuição de frequências para a produção (g/planta) de 25 plantas de uma progênie F2 de soja – DME/UFPEL

	Classes			f_i	fa_i	fr_i	far_i	PM_i
1	3,07	|---	8,440	3	3	0,12	0,12	5,755
2	8,44	|---	13,81	3	6	0,12	0,24	11,125
3	13,81	|---	19,18	8	14	0,32	0,56	16,495
4	19,18	|---	24,55	7	21	0,28	0,84	21,865
5	24,55	|---	29,92	1	22	0,04	0,88	27,235
6	29,92	|---|	35,30	3	25	0,12	1	32,61

b)

$$\overline{X} = \frac{[(\mathbf{3} * 5{,}755) + (\mathbf{3} * 11{,}125) + (\mathbf{8} * 16{,}495) + (\mathbf{7} * 21{,}865) + (1 * 27{,}235) + (3 * 32{,}61)]}{25}$$

$$= 18{,}4288$$

c)

$$\text{Md} = 13{,}81 + \frac{[(0{,}5 * 25) - 6)]}{8} * 5{,}37 = 18{,}1744$$

d)

A classe modal é a 3, pois é a de maior frequência simples.

e)

$\widehat{d}_i = \overline{X_i} - \overline{X}$	$\widehat{d}_i^{\,2}$	f_i	$\widehat{d}_i^{\,2} f_i$
-12,67	160,63	3,00	481,88
-7,30	53,35	3,00	160,04
-1,93	3,74	8,00	29,92
3,44	11,81	7,00	82,65
8,81	77,55	1,00	77,55
14,18	201,11	3,00	603,32
	soma	25,00	1435,35
		$\widehat{\sigma}^2$	57,41

BIBLIGRAFIA

- COSTA, S. F. **Introdução ilustrada à estatística**. 5.ed. São Paulo: Harbra, 2013. 399p.
- FERREIRA, D. F. **Estatística Básica**. Lavras: UFLA, 2005. 664p.

- MORETTIN, L. G. **Estatística básica**: probabilidade e inferência. São Paulo: Pearson Prentice Hall, 2010. 375p.
- REGAZZI, A. J. **Curso de iniciação a estatística**: roteiro de aulas da disciplina EST 105. Viçosa: UFV, 1997. 141p.
- RIBEIRO JÚNIOR, J. I. **Análises estatísticas no Excel**: guia prático. Viçosa: UFV, 2004. 251p.