Introdução à Estatística

A estatística é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.

A estatística é composta pelas seguintes áreas:

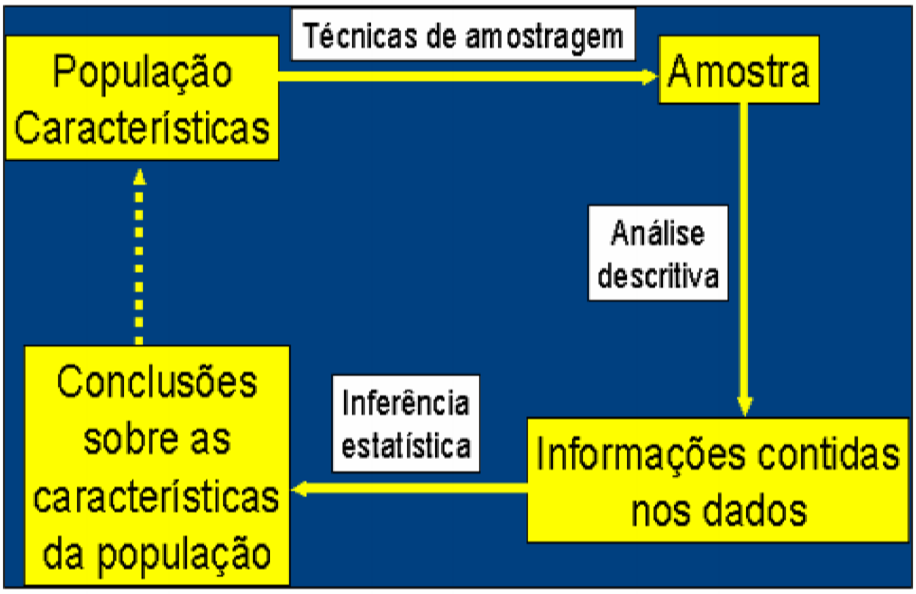
- Estatística Descritiva
- Probabilidade
- Inferência Estatística

A **estatística descritiva** é a etapa inicial da análise utilizada para descrever e resumir os dados. A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou está área da estatística.

A **teoria de probabilidades** nos permite descrever os fenômenos aleatórios, ou seja, aqueles em que está presente a incerteza.

A **inferência estatística** é o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados, das informações e conclusões obtidas a partir da amostra.

Etapas da Análise Estatística

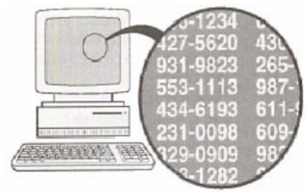
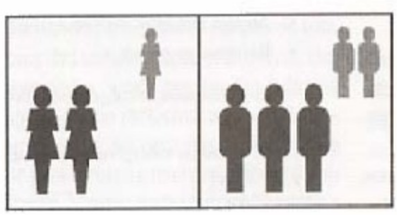
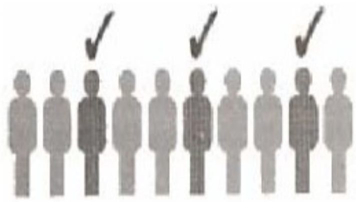
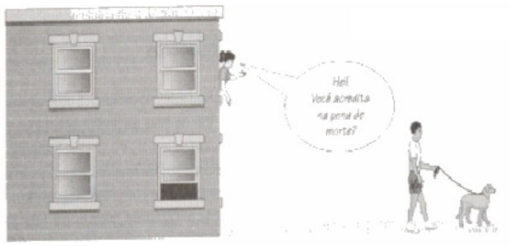


Amostragem

Uma área importante em muitas aplicações Estatísticas é a da Tecnologia de Amostragem.

Exemplos de Aplicação:

- Pesquisa de mercado,
- Pesquisa de opinião,
- Avaliação do processo de produção,
- Praticamente em todo experimento

Amostragem Aleatória Cada elemento da população tem a mesma chance de ser escolhido.	
Amostragem Estratificada Classificar a população em, ao menos dois estratos e extrair uma amostra de cada um.	
Amostragem Sistemática Escolher cada elemento de ordem k.	
Amostragem por Conglomerados Dividir em seções a área populacional, selecionar aleatoriamente algumas dessas seções e tomar todos os elementos das mesmas.	
Amostragem de Conveniência Utilizar resultados de fácil acesso	

Estatística Descritiva

O que se faz com as observações coletadas?

Resposta: 1ª etapa - resumo dos dados.

Variável

Qualquer característica associada a uma população. É a característica que vai ser observada, medida ou contada nos elementos da população ou da amostra e que pode variar, ou seja, assumir um valor diferente de elemento para elemento.

Não basta identificar a variável a ser trabalhada, é necessário fazer-se distinção entre os tipos de variáveis:

Classificação de variáveis		
Qualitativa	Nominal	sexo, cor dos olhos
	Ordinal	classe social, grau de instrução
Quantitativa	Contínua	peso, altura, salário
	Discreta	número de filhos, número de carros

Medidas Descritivas

Uma das maneiras de se resumir os dados de uma variável quantitativa, são usar o que chamamos de medidas descritivas que auxiliam a análise do comportamento dos dados. Tais dados são provenientes de uma população ou de uma amostra, o que exige uma notação específica para cada caso.

Classificam-se as medidas descritivas como: medidas posição (tendência central), medidas de dispersão, medidas de assimetria e de curtose.

Medidas de Tendência Central

As medidas de tendência central são assim denominadas por indicarem um ponto em torno do qual se concentram os dados. A seguir, são definidas as principais medidas de tendência central: média, mediana e moda.

Média aritmética (μ ou $\bar{x}$)

A média aritmética (X) é a soma de todos os valores observados da variável dividida pelo número total de observações. Sob uma visão geométrica a média de uma distribuição é o centro de gravidade, representa o ponto de equilíbrio de um conjunto de dados. É a medida de tendência central mais utilizada para representar a massa de dados.

Seja (x_1, ..., x_n) um conjunto de dados. A média é dada por:

$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

para dados populacionais ou amostrais, respectivamente.

Caso os dados estejam apresentados segundo uma distribuição de frequência, tem-se:

$$\mu=\frac{x_1F_1+x_2F_2+x_3F_3+...+x_nF_n}{n}=\frac{\sum_{i=1}^{n}x_iF_i}{n}$$

Observe que no caso de dados agrupados a média é obtida a partir de uma ponderação, onde os pesos são as frequências absolutas de cada classe e xi é o ponto médio da classe i.

Citam-se a seguir, algumas propriedades da média aritmética:

1. a média é um valor calculado facilmente e depende de todas as observações
2. é única em um conjunto de dados e nem sempre tem existência real, ou seja, nem sempre é igual a um determinado valor observado
3. a média é afetada por valores extremos observados
4. por depender de todos os valores observados, qualquer modificação nos dados fará com que a média fique alterada. Isto quer dizer que somando-se, subtraindo-se, multiplicando-se ou dividindo-se uma constante a cada valor observado, a média ficará acrescida, diminuída, multiplicada ou dividida desse valor.
5. a soma da diferença de cada valor observado em relação à média é zero, ou seja, a soma dos desvios é zero. $\sum(x_i-\bar{x})=0$

A propriedade 5, é de extrema importância para a definição de variância, uma medida de dispersão a ser definida posteriormente.

Destaca-se, ainda, que a propriedade 3, quando se observam no conjunto dados discrepantes, faz da média uma medida não apropriada para representar os dados. Neste caso, não existe uma regra prática para a escolha de uma outra medida.

Exemplo:

Para ilustrar, considere o número de filhos, por família, para um grupo de 8 famílias: 0, 1, 1, 2, 2, 2, 3, 4. Neste caso, a média é μ= 1,875 = filhos por família.

Entretanto, incluindo ao grupo uma nova família com 10 filhos, a média passa a ser μ= 2,788, o que eleva em 48,16% o número médio de filhos por família.

Assim, ao observar a média, pode-se pensar que a maior parte das famílias deste grupo tem três filhos quando, na verdade, apenas uma tem três filhos.

Moda (M_o)

A moda (M_o) é o valor que apresenta a maior frequência da variável entre os valores observados. Para o caso de valores individuais, a moda pode ser determinada imediatamente observando-se o rol ou a frequência absoluta dos dados.

É relevante salientar que um conjunto de dados pode apresentar todos seus elementos com a mesma frequência absoluta, e neste caso não existirá um valor modal, o que significa que a distribuição será classificada como amodal.

Pode ocorrer, também, casos em que a sequência de observações apresente vários elementos com frequências iguais, implicando numa distribuição plurimodal.

O uso da moda é mais indicado quando se deseja obter, rapidamente, uma medida de tendência central. Um outro aspecto que favorece a utilização da moda é que seu valor não é afetado pelos valores extremos do conjunto de dados analisado.

Exemplo:

Seja um grupo de pessoas com idades de 2, 3, 1, 2 e 50 anos. Temos que $M_o = 2$.

A Mo = 2 demostra maior eficiência para caracterizar o grupo que a média aritmética.

Mediana (M_e)

A mediana (M_e) é o valor que ocupa a posição central da série de observações de uma variável, em rol, dividindo o conjunto em duas partes iguais, ou seja, a quantidade de valores inferiores à mediana é igual à quantidade de valores superiores a mesma.

Exemplo:

Retomando o exemplo do número de filhos por famílias, verifica-se que: para o caso de oito famílias, **n=8 (par)**, a mediana é determinada como a seguir:

x	x_1	x_2	x_3	x_4		x_5	x_6	x_7	x_8
Valor Esperado	0	1	1	2	$\frac{x_4+x_5}{2}$	2	2	3	4
	4 observações				M_e=2	4 observações			

Quando se acrescenta ao grupo uma outra família com 10 filhos o tamanho da amostra passa a ser **n=9 (ímpar)**. Neste caso, a mediana é:

x	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Valor Esperado	0	1	1	2	2	2	3	4	10
	4 observações				M_e=2	4 observações			

Observe que nos dois casos, por coincidência, a mediana manteve-se, M_e=2, significando que 50% das famílias possuem menos de 2 filhos ou 50% possuem mais de 2 filhos. Mostra-se assim, que a mediana não é influenciada por valores extremos. Este procedimento pode tornar-se inadequado quando o conjunto de dados for composto por muitos elementos.

Medidas de dispersão

As medidas de dispersão auxiliam as medidas de tendência central a descrever o conjunto de dados adequadamente e indicam se os dados estão, ou não, próximos uns dos outros.

Desta forma, não há sentido calcular a média de um conjunto onde **não há variação** dos seus elementos. Existe ausência de dispersão e a medida de dispersão é igual a zero. Por outro lado, aumentando-se a dispersão, o valor da medida aumenta e se a variação for muito grande, a média não será uma medida de tendência central representativa.

Faz-se necessário, portanto, ao menos uma medida de tendência central e uma medida de dispersão para descrever um conjunto de dados.

As três medidas de dispersão que serão definidas a seguir, são:

- amplitude total
- variância
- desvio padrão

Amplitude Total (At)

A amplitude total de um conjunto de dados é a diferença entre o maior e o menor valor observado. A medida de dispersão não levar em consideração os valores intermediários perdendo a informação de como os dados estão distribuídos e/ou concentrados.

$$At = x_{max} - x_{min}$$

Exemplo: A amplitude total da idade dos alunos que cursam essa disciplina é: At = 37-18 = 19 anos, isto é, as idades dos alunos diferem em 19 anos.

Desvio-médio (D_m)

A diferença entre cada valor observado e a média é denominado desvio.

Ao somar todos os desvios, ou seja, ao somar todas as diferenças de cada valor observado em relação a média, o resultado é igual a zero (propriedade 5 da média). Isto significa que esta medida não mede a variabilidade dos dados. Para resolver este problema, pode-se desconsiderar o sinal da diferença, considerando-as em módulo e a média destas diferenças em módulo é denominada desvio médio:

$$D_m = \frac{\sum_{i=1}^{n} |x_i - \mu|}{n}$$

Variância (δ^2) e Desvio Padrão (δ)

Enquanto não há nada conceitualmente errado em se considerar o desvio médio, esta medida não tem certas propriedades importantes e não é muito utilizada. O mais comum é considerar o quadrado dos desvios em relação à média e então calcular a média. Obtém-se, assim a variância que é definida por:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

Entretanto, ao calcular a variância observa-se que o resultado será dado em unidades quadráticas, o que dificulta a sua interpretação. O problema é resolvido extraindo-se a raiz quadrada da variância, definindo-se, assim, o desvio padrão:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}$$

Exemplo:

Considere três alunos cujas notas em uma disciplina estão apresentadas na tabela abaixo. Observa-se que as médias das notas dos três alunos são iguais, porém, seus desvios em torno da média são diferentes. Isto quer dizer que seus desempenhos são diferentes. O aluno A é constante em seu desempenho, o segundo vai progredindo aos poucos e o terceiro diminui abruptamente seu desempenho. Em outras palavras, apesar dos três alunos terem o mesmo desempenho médio, a variabilidade difere.

Aluno	Notas	Soma	Média μ	$d=x_i-\mu$	$\lvert x_i-\mu \rvert$	$(x_i-\mu)^2$	$\sqrt{\sum (x_i-\mu)^2}$
A	8	40	8	0	0	0	$\sqrt{0}=0$
	8			0	0	0	
	8			0	0	0	
	8			0	0	0	
	8			0	0	0	
Total				0	0	0	
B	6	40	8	-2	2	4	$\sqrt{16}=4$
	6			-2	2	4	
	8			0	0	0	
	10			2	2	4	
	10			2	2	4	
Total				0	8	16	
C	10	40	8	2	2	4	$\sqrt{30}=5{,}48$
	10			2	2	4	
	10			2	2	4	
	5			-3	3	9	
	5			-3	3	9	
Total				0	12	30	

Como demonstrado no exemplo, geralmente, o desvio padrão é maior ou igual ao desvio médio, e isto devido ao fato de que para o cálculo do desvio-padrão, cada desvio em torno da média é elevado ao quadrado, aumentando desproporcionalmente o peso dos valores extremos.

Exercícios:

Em um treinamento de salto em altura, os atletas realizaram 4 saltos cada um. Veja as marcas obtidas por 3 atletas:

Atleta A	148 cm	170 cm	155 cm	131 cm
Atleta B	145 cm	151 cm	150 cm	152 cm
Atleta C	146 cm	151 cm	143 cm	160 cm

a) Qual deles obteve melhor média?
b) Qual deles foi o mais regular?

Referências Bibliográficas

BARBETTA, PEDRO ALBERTO; REIS, MARCELO MENEZES; BORNIA, ANTONIO CESAR. **Estatística para Cursos de Engenharia e Informática**. São Paulo: ed. Atlas, 2004

DANTE, LUIS ROBERTO. **Matemática Contexto e Aplicações, vol 2. 3ª ed.** São Paulo: ed. Ática, 2004

DANTE, LUIS ROBERTO. **Matemática Contexto e Aplicações, vol 3. 3ª ed.** São Paulo: ed. Ática, 2004

Probabilidade e Estatística. Disponível em: <https://www.youtube.com/watch?v=eFyAyz6Xy6g> acesso em 08 fev 2018