Estatística Descritiva

cognitiza

Your Data & AI school

Prof. Altino Dantas

“

Estatística é a arte de torturar os números até que eles confessem

— Piada no meio acadêmico

“

A estatística é uma parte da Matemática que fornece métodos para a **coleta**, **organização**, **descrição**, **análise** e **interpretação** de dados, viabilizando a utilização deles na **tomada de decisões**.

Descritiva

Inferencial

Probabilidade

Cultura

6 razões para acreditar que estatística é a profissão do futuro

Por **Da Redação** Atualizado em 31 out 2016, 19h05 - Publicado em 15 abr 2011, 22h00

1. Dá para trabalhar onde você quiser
2. O próximo Einstein será um estatístico
3. Sobra informação
4. Falta gente
5. Sobrevive a crises
6. Todo mundo entende

Vamos ao trabalho!

Estatística Descritiva

Estatística descritiva

Conjunto de **métodos** e **procedimentos** para a apresentação sumarizadas das características de dados amostrais, através de **tabelas** ou **gráficos**;

Apresenta, em suma, medidas de posição, tendência central e medidas de dispersão.

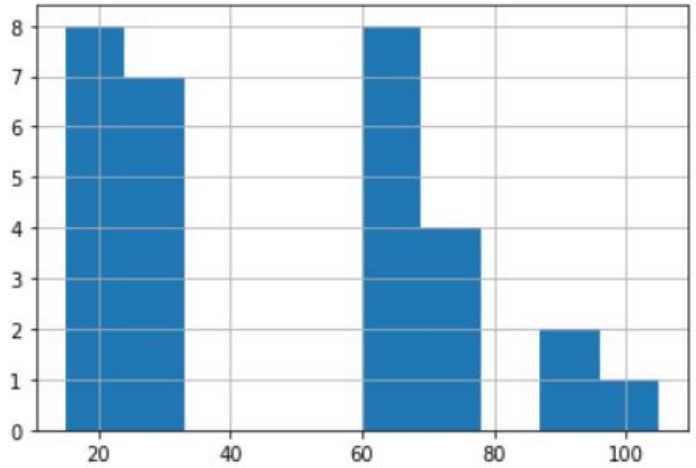
Visualizando os dados de uma amostra

- **Distribuição de Frequência** - dados agrupados por classes ou categorias contabilizado-se as ocorrências de cada classe;
- **Histograma** - forma gráfica de apresentar uma distribuição de frequência;
- **Outros gráficos** - gráfico de torta ou polígono de frequências.

Entregas app de delivery

	Produto	dia-da-semana	hora-do-pedido	tempo-entrega	tempo-entrega-real
0	A	1	8	15	15
1	A	6	17	15	60
2	B	6	14	30	30
3	A	1	14	45	15
4	B	2	1	60	75
5	B	1	13	30	30
6	B	2	14	75	15
7	B	7	1	15	60
8	D	4	9	15	15
9	X	5	20	30	30
10	C	4	22	30	15
11	C	4	5	60	60
12	C	3	1	45	75

Figura 1.



Tempo de entrega real

...

Medidas de posição ou tendência central

- Medida de posição ou percentis:
  - Máximo
  - Mínimo
  - Quartis
  - Decis
- Medidas de tendência central, definem o centro da distribuição:
  - Média
  - Mediana
  - Moda

Medidas de posição

Considere os dados abaixo com uma amostra para uma variável **X** qualquer

X	5	20	5	5	10	15	60	45	45	90
i	1	2	3	4	5	6	7	8	9	10

Medidas de posição

- Máximo: maior valor observado;
- Mínimo: menor valor observado;
- Quartis: dividem a amostra em 4 partes iguais, ou seja, três quartis:
  - Para uma amostra com **n** elementos, o primeiro quartil (Q1) será o elemento de ordem n/4, o segundo (Q2) o elemento 2n/4 e o terceiro (Q3) o elemento 3n/4
  - Se as ordens dos elementos referentes aos quartis resultarem em números fracionários, arredonde-os para cima;
  - Se as ordens resultarem em números inteiros, tome a média aritmética deste com o seguinte;

5	5	5	10	15	20	45	45	60	90

Q1 Q2 Q3

Média

Se x_1, ..., x_n são os n valores (distintos ou não) da variável ***X***, a média aritmética, ou simplesmente média, de X pode ser escrita:

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

X	5	20	5	5	10	15	60	45	45	90
i	1	2	3	4	5	6	7	8	9	10

$\overline{X} =$ Média = 30

Média

Se x_1, ..., x_n são os n valores (distintos ou não) da variável ***X***, a média aritmética, ou simplesmente média, de X pode ser escrita:

1

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

2

$$\bar{X} = \frac{n_1 X_1 + n_2 X_2 + \ldots + n_k X_k}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i X_i$$

n_1 são iguais a x_1, n_2 são iguais a x_2 etc., n_k iguais a x_k

3

$$\bar{X} = \sum_{i=1}^{k} f_i X_i$$

f_i = frequência relativa de x_i

Mediana

Considera as ***n*** observações de uma variável ***X*** ordenadas de forma não-decrescente (eventualmente, crescente) e é dada por:

$$\mathrm{md}(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ ímpar;} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ par.} \end{cases}$$

	5	5	5	10	15	20	45	45	60	90
i	1	2	3	4	5	6	7	8	9	10

Média = 17,5

Média vs Mediana

- Você pretende analisar o tempo que as pessoas permanecem em ligações do *call center* da sua empresa quando estão interessadas no assunto **“cancelar assinatura”**. Para isso, você captura ligações ao longo de um dia e constata que muitas dessas ligações perduram por tempo similar, enquanto algumas duram muito mais tempo.
- Para verificar a tendência central dos dados seria mais adequado usar média ou mediana? Qual a consequência do uso de cada uma delas?

Moda

- Valor que ocorre com maior freqüência
  - Pode haver mais de uma moda
  - Para o exemplo anterior, a moda é 5 (3 ocorrências)
- Basta verificar a distribuição de frequências

5	20	5	5	10	15	60	45	45	90

Moda = 5

Observações

- Para calcular a **moda** de uma variável, precisamos apenas da distribuição de freqüências (contagem).
- Já para a mediana necessitamos minimamente ordenar as realizações da variável.
- Finalmente, a média só pode ser calculada para variáveis quantitativas.

Medidas de dispersão

- O **resumo** de um conjunto de dados por uma única medida representativa de posição central **esconde** toda a informação sobre a **variabilidade** do conjunto de observações;
- As medidas de dispersão medem a dispersão ou espalhamento (absoluto ou relativo) dos elementos de uma amostra.

grupo A (variável X): 3, 4, 5, 6, 7
grupo B (variável Y): 1, 3, 5, 7, 9
grupo C (variável Z): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

$$\bar{x} = \bar{y} = \bar{z} = \bar{w} = \bar{v} = 5{,}0$$

Medidas de dispersão

grupo A (variável X): 3, 4, 5, 6, 7
grupo B (variável Y): 1, 3, 5, 7, 9
grupo C (variável Z): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

$$\bar{X} = \bar{y} = \bar{z} = \bar{W} = \bar{v} = 5{,}0$$

$$\mathrm{dm}(X) = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n},$$

Desvio médio

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n},$$

Variância

Medidas de dispersão

grupo A (variável X): 3, 4, 5, 6, 7
grupo B (variável Y): 1, 3, 5, 7, 9
grupo C (variável Z): 5, 5, 5, 5, 5
grupo D (variável W): 3, 5, 5, 7
grupo E (variável V): 3, 5, 5, 6, 6

$$\mathrm{dm}(X) = \frac{\sum_{i=1}^{n} |X_i - \overline{X}|}{n},$$

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n},$$

Grupo A

$$\mathrm{dm}(X) = 6/5 = 1{,}2,$$
$$\mathrm{var}(X) = 10/5 = 2{,}0,$$

Grupo D

$$\mathrm{dm}(W) = 4/4 = 1{,}0,$$
$$\mathrm{var}(W) = 8/4 = 2{,}0.$$

Pelo desvio médio, o Grupo D é mais homogêneo do que o A

Medidas de dispersão

- Sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em *cm*, a variância será expressa em cm^2);
- Para contornar essa questão, costuma-se usar o desvio padrão, que é a raiz quadrada da variância:

$$\mathrm{dp}(X) = \sqrt{\mathrm{var}\ (X)}$$

Desvio-padrão como medida de risco

Exemplo – considere dois fundos de investimento, A e B, e seus respectivos retornos nos últimos 12 meses (em %)

	jan	fev	mar	abr	mai	jun	jul	ago	set	out	nov	dez
Fundo A	5,25	4,37	5,36	5,64	4,77	4,51	4,82	4,93	5,39	5,29	4,68	5,04
Fundo B	1,23	2,28	3,40	7,00	6,75	7,47	7,28	5,29	10,7	5,37	-0,2	3,45

Em qual dos dois fundos você colocaria o seu dinheiro?

Exercício

- Crie histograma e piechart no google sheets ou colab para a variável dia da semana dos dados da Figura 1;
- Considerando os dados da Figura 1, determine os quartis para a variável hora-do-pedido. Use o google Sheets para fazê-lo.
- Crie uma tabela com as frequências relativas dos dados do slide 11 e mostre que as fórmulas 1, 2 e 3 são equivalentes.
- Crie um gráfico de dispersão para as variáveis tempo-entrega e tempo-entrega-real (google sheets), da Figura 1;

Obrigado.