1

Dicionário / definições

População: conjunto de todos os elementos de interesse em estudo

Amostra: subconjunto representativo da população que será estudado para tirar conclusões para a população toda

Variável: toda característica que, observada em uma unidade experimental, pode variar de uma unidade para outra

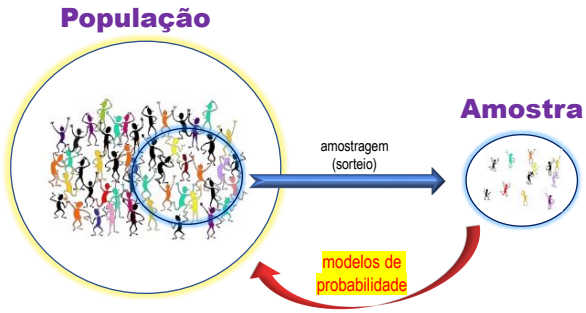
Parâmetro: medida que descreve alguma característica da população

Estimador ou estatística: medida que descreve alguma característica da amostra

2

População x Amostra: etapas de uma análise de dados

Para inferir (deduzir) certas características de uma população (pessoas entrevistadas, peças, repetições de um processo, etc...) deve-se trabalhar com uma amostra que seja representativa dessa população.



Estatística descritiva
- Consistência dos dados
- Interpretações iniciais

Inferência estatística
- Estimação de quantidades desconhecidas
- Extrapolação dos resultados

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ

3

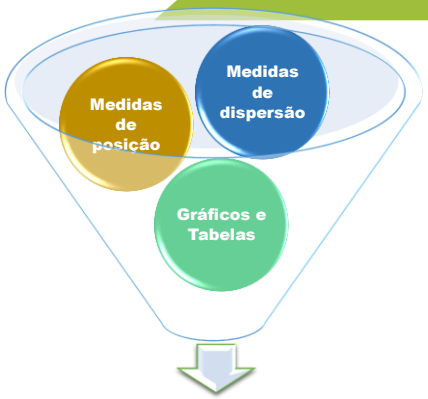
Estatística descritiva

- ✓ Trata-se da organização, apresentação e descrição de um conjunto de dados (para uma ou mais variáveis);
- ✓ Os resumos descritivos podem ser organizados em tabelas, apresentados graficamente ou a partir de estimadores ou estatísticas de parâmetros da população.

Gráficos e Tabelas — Resumos visuais

Medidas de posição — São valores calculados que tendem a representar melhor a distribuição de um conjunto de dados em um valor central

Medidas de dispersão — São medidas que indicam a dispersão dos dados em torno de algum valor central



INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ

4

Classificação das variáveis

Saber classificar cada tipo de variável auxilia na busca de técnicas estatísticas mais adequadas para o resumo dos dados.

Qualitativa
Exs.: Sexo, Classe social, Qualidade da peça (boa ou defeituosa), Escolaridade, ...

Quantitativa
Exs.: Salário (R$), Nº defeitos por peça, Volume (mL), Estatura (m), Nº filhos,...

- **Discreta**: Ex.: Nº defeitos por peça, Nº de filhos...
- **Contínua**: Ex.: Salário (R$), Peso (kg), Estatura (m),...

Qualitativa: as respostas desse tipo de variável representam diferentes categorias que se distinguem por alguma característica não numérica.

Quantitativa: as respostas desse tipo de variável consistem em números que representam, em geral, contagem (discretas) ou medidas provenientes de alguma mensuração (contínuas).

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ

5

Exercício: Classifique as variáveis do exemplo abaixo.

De acordo com a Organização Mundial da Saúde, o AVC é a 2ª causa de morte no mundo, responsável por aproximadamente 11% do total de mortes.

O conjunto de dados AVC é usado para prever se um paciente tem probabilidade de desenvolver AVC com base em parâmetros de entrada como sexo, idade, várias doenças e tabagismo. Cada linha nos dados fornece informações relevantes sobre o paciente.

Na planilha AVC do arquivo "Aula02.xlsx" são apresentadas as informações de 5110 pessoas (parte do arquivo é apresentado no quadro abaixo):

* hipertensão, doença cardíaca e avc: 0 – não e 1 - sim

id	gender	age	hypertension*	heart_disease*	ever_married	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke*
9046	Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202,21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174,12	24	never smoked	1
⋮											

6

Classificação das variáveis?

id	gender	age	hypertension	heart_disease	ever_married	work_type	residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202,21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174,12	24	never smoked	1
						⋮					
44873	Female	81	0	0	Yes	Self-employed	Urban	125,2	40	never smoked	0
19723	Female	35	0	0	Yes	Self-employed	Rural	82,99	30,6	never smoked	0
37544	Male	51	0	0	Yes	Private	Rural	166,29	25,6	formerly smoked	0
44679	Female	44	0	0	Yes	Govt_job	Urban	85,28	26,2	Unknown	0
?	?	?	?	?	?	?	?	?	?	?	?

7

Classificação das variáveis

Para cada cada tipo de variável existem técnicas estatísticas mais adequadas para o resumo dos dados.

Qualitativa

Exs.: Sexo, Classe social, Qualidade da peça (boa ou defeituosa), Escolaridade, ...

Resumos estatísticos que podem ser feitos

- ✓ tabelas com a **frequência absoluta** de cada categoria
- ✓ tabelas com a **frequência relativa (%)** de cada categoria
- ✓ construir gráficos de pizza, coluna, barras, ...

Quantitativa

Exs.: Salário (R$), Nº defeitos por peça, Volume (mL), Estatura (m), Nº filhos,...

Resumos estatísticos que podem ser feitos

- ✓ cálculo de medidas de posição (ou de localização)
- ✓ cálculo de medidas de dispersão (ou variabilidade)
- ✓ construir gráficos: boxplot, histograma, linha, dispersão, ...

8

Reflexão: por quê calcular média? É algo tão simplório...

9

Medidas de posição

Estatísticas que tendem a representar melhor a distribuição dos dados de uma variável X em um único valor central. Fornecem uma ideia do "centro de gravidade" dos dados.

✓ Média da amostra ($\bar{x}$)

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

x_i : valor da i-ésima observação da variável X
n : tamanho da amostra

✓ Mediana (Md)

É o valor que ocupa a posição central quando os dados estão ordenados

$$Md = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, \text{se n for ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}}{2}, \text{se n for par} \end{cases}$$

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq x_{(4)} \leq x_{(5)} \leq \ldots \leq x_{(n)}$$

menor valor da variável X

maior valor da variável X

✓ Moda (Mo)

É o valor (ou valores) de maior frequência na amostra (OBS.: não tão usada na prática)

10

Medidas de dispersão (ou variabilidade)

São medidas que indicam a dispersão dos dados em torno de algum valor central

✓ **Variância amostral (s^2)**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \text{ ou } = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

✓ **Desvio padrão (s)**

O desvio padrão é definido pela raiz quadrada positiva da variância: $s = \sqrt{s^2}$

✓ **Coeficiente de variação ($CV_{\%}$)**

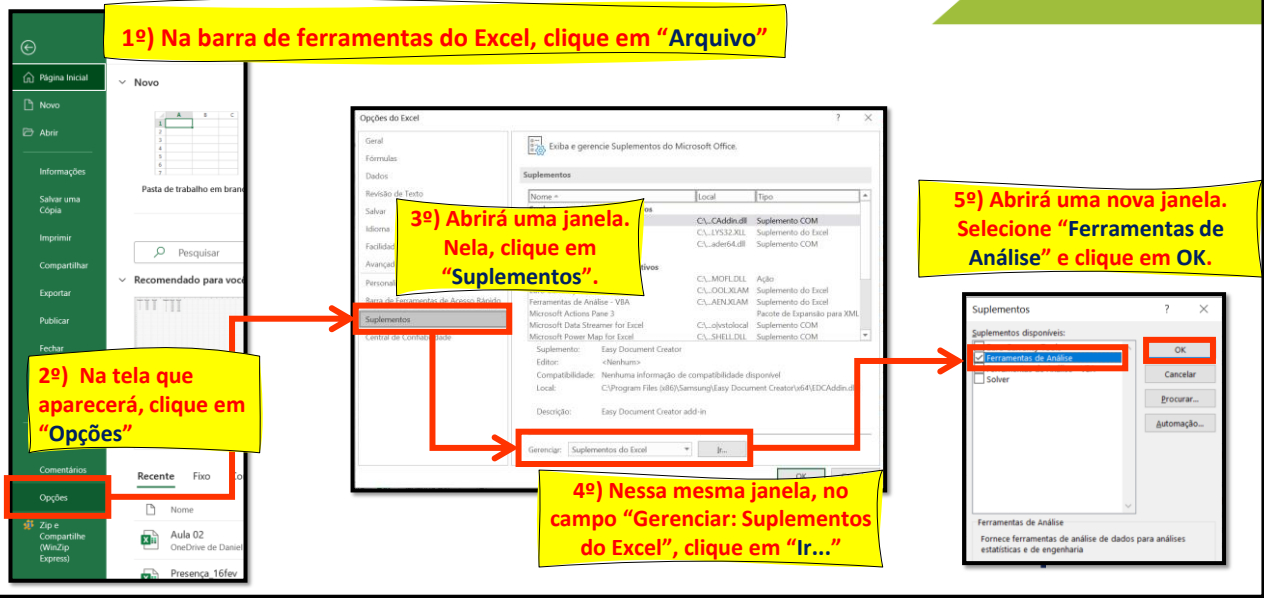
Indica a dispersão de um conjunto de dados em relação à sua média

Não existe um consenso, mas na prática, uma variável com $CV_{\%}$ superior a 30% (ou 40%) é considerada como tendo alta variabilidade

$$CV_{\%} = \frac{s}{\bar{x}} \cdot 100$$

11

Como habilitar a Análise de Dados no Excel (Windows)



12

Exercício. Excel: Dados → Análise de dados

De acordo com a Organização Mundial da Saúde, o AVC é a 2ª causa de morte no mundo, responsável por aproximadamente 11% do total de mortes.

O conjunto de dados AVC é usado para prever se um paciente tem probabilidade de desenvolver AVC com base em parâmetros de entrada como sexo, idade, várias doenças e tabagismo. Cada linha nos dados fornece informações relevantes sobre o paciente.

Na planilha AVC do arquivo "Aula02.xlsx" são apresentadas as informações de 5110 pessoas.

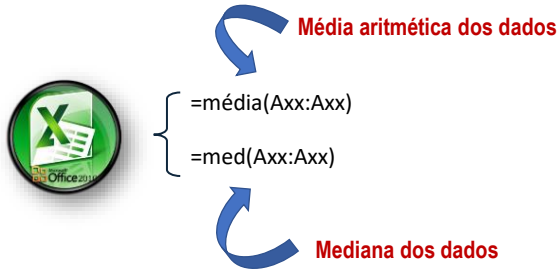
a) Calcule resumos numéricos univariados usando a ferramenta de Análise de Dados do Excel.

b) Quais são os prós e contras dessa ferramenta?

13

Média x Mediana

Na planilha bônus do arquivo "Aula02.xlsx" são apresentados quatro cenários (A, B, C e D) de valores de bônus de Natal pagos a uma amostra de estagiários. Calcule a média e mediana da variável "bônus de final de ano de estagiário" e compare os resultados dessas duas medidas de posição para cada cenário.



INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ

14

Por que medir a variabilidade de uma variável?

Na planilha bônus do arquivo "Aula02.xlsx" são apresentados quatro cenários (A, B, C e D) de valores de bônus de Natal pagos a uma amostra de estagiários. Calcule o desvio padrão e o coeficiente de variação e avalie em quais situações a variabilidade dos dados é maior.

Variância dos dados amostrais

=var.a(Axx:Axx)

=desvpad.a(Axx:Axx)

Desvio padrão dos dados amostrais

OBSERVAÇÕES

- Com as fórmulas acima, obtemos a variância e o desvio padrão **amostrais** de uma variável de interesse;
- O Excel oferece calcular a variância e o desvio padrão populacionais (*var.p* e *desvpad.p*, respectivamente). Na prática, não são muito utilizadas;
- Não existe uma fórmula pronta no Excel para calcular o CV.

15

Média e Mediana com Distribuição de Frequências

Quando os dados estão dispostos em uma tabela de frequências com k classes, calcula-se a média amostral pela expressão

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n}$$

não tem fórmula pronta no Excel!!!

Número de defeitos	Freqüência
0	13
1	11
2	9
3	8
4	6
5	1
6	2

7 classes

Exercício. Uma amostra de 50 peças foi selecionada pelo controle de qualidade de uma empresa. A variável X de interesse é o número de defeitos por peça. Em média, quantos defeitos por peça há nessa amostra?

$\bar{x} \cong 1{,}9$ defeitos

Pense em como calcular a mediana nesse caso. Quanto ela vale?

Md = 2 defeitos

16

Variância e DP com uma Distribuição de Frequências

Numa tabela de frequências, composta de k classes, a variância amostral pode ser calculada por:

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i}{n-1} \text{ ou } = \frac{1}{n-1}\left[\sum_{i=1}^{k} x_i^2 f_i - \frac{\left(\sum_{i=1}^{k} x_i f_i\right)^2}{n}\right]$$

não tem fórmula pronta no Excel!!!

7 classes

Número de defeitos	Freqüência
0	13
1	11
2	9
3	8
4	6
5	1
6	2

Exercício. Uma amostra de 50 peças foi selecionada pelo controle de qualidade de uma empresa. A variável X de interesse é o número de defeitos por peça. Quanto vale o desvio padrão de X?

$s \cong 1{,}7$ defeitos

17

Distribuição de Frequências com Dados Agrupados

Quando os dados estão dispostos em uma tabela de frequências com k classes, porém com valores agrupados, utiliza-se as mesmas expressões apresentadas nos dois *slides* anteriores.

Rendimento mensal (%)	Freqüência
0,60 ⊢ 0,70	4
0,70 ⊢ 0,80	2
0,80 ⊢ 0,90	4
0,90 ⊢ 1,00	2

Nesse caso, cada x_i representa o ponto médio da classe i

Exercício Uma amostra do rendimentos mensais de certa aplicação financeira foi selecionada e os dados foram apresentados consolidados numa tabela de frequência agrupada. Quais os valores da média e do desvio padrão do retorno financeiro? Calcule.

$\bar{x} \cong 0{,}783\ \%$

$s^2 \cong 0{,}013\ \%^2$ → $s \cong 0{,}115\ \%$

Como calcular a média e a variância da amostra agora?

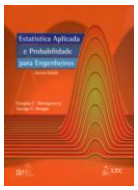
18

Leitura e exercícios recomendados

- Fazer eventuais exercícios não finalizados na aula

Cap. 1

Seção 1.1 a 1.5 e seus respectivos exercícios



Cap. 1 e Cap. 6

Seção 6.1 e seus respectivos exercícios

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ

19

INSTITUTO MAUÁ DE TECNOLOGIA
MAUÁ

Campus São Caetano do Sul
Praça Mauá, 01 - São Caetano do Sul - SP

20