



Linguagem R

Aula 8

- Objetivo da aula:
- Amostragem simples
- Amostragem sistemática
- Amostragem estratificada
- Medidas de centralidade e variabilidade

Estatística: Conceitos Básicos.

Estatística é conjunto de técnicas/métodos para a coleta, organização, análise e interpretação de dados.

Dado é o valor assumido por uma variável aleatória em determinado experimento.

População é um conjunto com todos os dados.

Amostra é um subconjunto da população.

Evento é cada resultado possível em um experimento.

Estatística: Conceitos Básicos.

Variáveis Qualitativas são variáveis que indicam qualidades, atributos, características não numéricas de forma geral.

Variáveis Quantitativas são variáveis que indicam medidas, contagens, etc.

Frequência representa o número de vezes que um valor ocorre em um conjunto de dados.

Amostragem Simples:

Uma **amostragem aleatória simples** dá então a cada elemento do público-alvo a mesma probabilidade de serem selecionados, visto que essa seleção é feita em forma de sorteio.

Exemplificando Amostragem Simples:

Um farmacêutico mistura bem um recipiente com 1000 comprimidos de paracetamol e retira, então, 50 comprimidos que devem ser testados para verificar o conteúdo exato.

Amostragem Sistemática:

A **amostragem sistemática** é um tipo de amostragem probabilística, onde se faz uma seleção aleatória do primeiro elemento para a amostra e logo se selecionam os itens subsequentes utilizando intervalos fixos ou sistemáticos até se chegar ao tamanho da amostra desejada.

Exemplificando Amostragem Sistemática:

Um engenheiro de controle da qualidade seleciona cada centésima fonte de computador que passa em uma esteira transportadora.

Amostragem Estratificada:

A **amostragem estratificada** é um método de amostragem que envolve a divisão de uma população em subgrupos menores, conhecidos como estratos.

Por sua vez, esses subgrupos (ou estratos) são formados a partir dos atributos ou características compartilhadas dos membros.

Exemplificando Amostragem Estratificada:

Um médico seleciona casos de diabetes conforme idade e nível de glicose no sangue.

Amostragem Estratificada:

Por exemplo: Um analista deseja pesquisar o número de estudantes de TI que receberam uma oferta de emprego um mês depois de formado num dado ano. O analista descobre que havia quase 10.000 formados em TI naquele ano.

Para fazer uma amostragem estratificada o analista deve criar grupos populacionais com base na idade, região, experiência.

Para isto ele deve ter noção de peso de cada estrato, por exemplo, idade(10%), região(60%), experiência(30%).

A partir daí o analista deve selecionar nos estratos amostras proporcionais a estes valores para que a pesquisa seja realizada da forma correta.

Exemplificando os tipos de amostragem:

Amostragem Simples:

10% dos alunos de uma população com notas entre 9 e 10 serão sorteados para receber uma bolsa de estudos em TI.

Exemplificando os tipos de amostragem:

Amostragem Sistemática:

Uma amostra de 10% dos alunos com déficit de atenção diagnosticado é selecionada.

Sorteia-se um valor de 1 a 5. Se o sorteado for o 2, incluem-se na amostra 2, o 7, o 12 e assim por diante de cinco em cinco.

Exemplificando os tipos de amostragem:

Amostragem Estratificada:

Supondo que dos 100 alunos de uma escola, 60 sejam meninos e 40 sejam meninas. Vamos obter a amostra proporcional estratificada de 10% desta população, ou seja, temos dois estratos: sexo masculino, sendo 6 meninos e sexo feminino, sendo 4 meninas.

Medidas de Posição:

Média aritmética é calculada somando-se todas as observações e dividindo o resultado pelo número de elementos foram somados.

Mediana é o elemento que ocupa a posição central do conjunto de dados.

Moda é o valor mais frequente no conjunto de dados.

Em R usa-se:

`mean()`, `median()`

Moda:

R não tem uma função *built-in* para determinar a moda. Então constrói-se uma função para se determinar a moda.

```
moda <- function(v) {  
  valor_unico <- unique(v)  
  valor_unico[which.max(tabulate(match(v, unique)))]  
}  
valor_moda <- moda(nome do DataFrame$nome da coluna)
```

Exemplo:

Para o conjunto de dados a seguir, determine o valor médio, a mediana e a moda.

conj = (2,3,3,4,4,4,5,5,5,5)

```
19 conj <- c(2,3,3,4,4,4,5,5,5,5)
20 mean(conj)
21 median(conj)
22 moda <- function(v) {
23   valor_unico <- unique(v)
24   valor_unico[which.max(tabulate(match(v, valor_unico)))]
25 }
26 valor_moda <- moda(conj)
27 print(valor_moda)
28
29
30
```

19:1


(Top Level) ↕

R Scrip

Console

Terminal ×

Background Jobs ×

 R 4.2.2 · C:/Users/qualquer/Downloads/arq1/ ↗

```
> conj <- c(2,3,3,4,4,4,5,5,5,5)
> mean(conj)
[1] 4
> median(conj)
[1] 4
> moda <- function(v) {
+   valor_unico <- unique(v)
+   valor_unico[which.max(tabulate(match(v, valor_unico)))]
+ }
> valor_moda <- moda(conj)
> print(valor_moda)
[1] 5
> |
```

Medidas de Dispersão:

A variância e o desvio padrão são as duas medidas de dispersão mais usadas em estatística.

O mais comum em ciência de dados é o uso do desvio padrão uma vez que este está na mesma unidade de medida do valor médio.

Em R, usa-se `sd()` para se determinar o desvio padrão e `var()` para se determinar a variância.

Use o exemplo anterior para determinar o desvio padrão e a variância.

Exemplo:

Para o conjunto de dados a seguir, determine o valor médio, a variância e o desvio padrão.

conj = (2,3,3,4,4,4,5,5,5,5)

Aplicando Estatística em um DataFrame:

Use o dataframe vendas.csv para o cálculo do valor médio, desvio padrão e variância.

Para isto faça:

```
setwd('o path do arquivo')
```

```
getwd()
```

```
vendas <- read.csv("vendas.csv", fileEncoding = "windows-1252")
```

```
str(vendas)
```

Aplicando Estatística em um DataFrame:

```
val_med <- mean(vendas$Valor)  
print(val_med)
```

```
desv_pad <- sd(vendas$Valor)  
print(desv_pad)
```

```
vari <- var(vendas$Valor)  
print(vari)
```

```
137 setwd('C:/Users/qualquer/Downloads/arq1')
138 getwd()
139 vendas <- read.csv("vendas.csv", fileEncoding = "windows-1252")
140 str(vendas)
141 val_med <- mean(vendas$valor)
142 print(val_med)
143 desv_pad <- sd(vendas$valor)
144 print(desv_pad)
145 vari <- var(vendas$valor)
146 print(vari)
147
```

145:1 (Top Level) ⬆

Console

Terminal ×

Background Jobs ×

R 4.2.2 · C:/Users/qualquer/Downloads/arq1/ ↗

```
> val_med <- mean(vendas$valor)
> print(val_med)
[1] 80
>
> desv_pad <- sd(vendas$valor)
> print(desv_pad)
[1] 27.38613
>
> vari <- var(vendas$valor)
> print(vari)
[1] 750
```