****

**UFFS – Universidade Federal da Fronteira Sul**

**Campus Cerro Largo**

**PROJETO DE EXTENSÃO**

**Software R: capacitação em análise estatística de dados utilizando um software livre**

Fonte: https://www.r-project.org/

**Blog do projeto:**

https://softwarelivrer.wordpress.com/equipe/

**Equipe:**

Coordenadora: Profe. Iara Endruweit Battisti

(iara.battisti@uffs.edu.br)

Colaboradores: Profa. Denize Reis

Prof. Erikson Kaszubowski

Prof. Reneo Prediger

Profa. Tatiane Chassot

Bolsista: Djaina Rieger

**Exercícios:**

Crie uma situação de pesquisa, e os possíveis resultados. Estes devem ser representados em um banco de dados que deve também ser disponibilizado (entregue).

Neste Banco de dados devem constar variáveis quantitativas discretas e contínuas e variáveis qualitativas.

Na sequência, UTILIZE pelo menos uma vez cada função estudada neste curso para descrever as variáveis que você criou no estudo fictício.

Data de entrega: 11/10/2017

**MÓDULO 2 – ESTATÍSTICA DESCRITIVA**

**Exemplo da aula**

Estaremos utilizando como exemplo os dados de uma pesquisa (dados simulados), cujo banco de dados está intitulado “Dados\_pesquisa.ods”. Os dados são referentes aos resultados obtidos por ocasião de uma pesquisa realizada entre os consumidores a fim de analisar características associadas ao mercado consumidor de sucos, sendo que a amostra é composta de 348 entrevistados aleatoriamente selecionados.

* O objetivo primário do estudo foi determinar variáveis que seriam úteis para caracterizar os consumidores que já conhecem o suco e a possibilidade potencial de futuros consumidores. Há também interesse nas relações entre variáveis das características pessoais desses consumidores ou futuros consumidores.
* A pesquisa foi realizada, depois que os participantes realizaram uma visita técnica às instalações da empresa e puderam conhecer seus produtos e processos.
* Para cada entrevistado foram registrados dados para as seguintes variáveis:

**Sexo** – Gênero sexual

**Divulgação** – Forma de acesso ao suco ou publicidade do mesmo

**Renda\_h** – Renda por hora do entrevistado

**Praticidade** – Aspectos quanto a oferta do suco, como por ex. embalagem

**Sabor** – Aspectos relacionados ao sabor

**Pessoas\_familia** – Número de pessoas que compõe o grupo familiar

**Preço** – como cada entrevistado classificava o preço do produto

**consumo\_anterior** – Se já consumia o suco antes da visita técnica

**consumo\_pos** – Se consumia o suco após a visita técnica

Pede-se:

1. Salvar inicialmente os dados em formato CSV.
2. Ler os dados no `Environment' pelo 'Import Dataset...From CSV'.
3. Carregar o banco de dados, com a finalidade de usar os objetos (variáveis) diretamente nas funções a serem utilizadas.

>attach(nome\_da\_planilha)

**Estatísticas descritivas**

1. Para determinar o valor máximo de um conjunto de dados, utilizamos:

>max(nome\_da\_variável)

Use a variável **Renda\_h**

1. De forma análoga, para determinar o valor mínimo de um conjunto de dados, utilizamos:

>min(nome\_da\_variável)

Use a variável **Renda\_h**

Obs. Para determinar a amplitude total de um conjunto de dados, utilizamos:

>max(nome\_da\_variável)-min(nome\_da\_variável)

Use a variável **Renda\_h**

1. Para obter as medidas da estatística descritiva, no caso medidas de tendência central (mínimo, quartil 1, mediana, média, quartil 3, máximo):

> summary(nome\_da\_variável)

Ex. Use a variável **Renda\_h**

1. A moda é o valor que tem o maior número de ocorrências em um conjunto de dados.

O R não tem um padrão de função embutida para calcular a moda. Uma sugestão é a criação de uma função pelo usuário, que pode ser obtida, por exemplo por:

>subset(table(nome\_da\_variável),table(nome\_da\_variável)==max(table(nome\_da\_variável))).

Ex. Use a variável **Praticidade**

Ex. Use a variável quantitativa **Pessoas\_familia**

**Obs.** O primeiro valor encontrado, refere-se ao valor da moda ao passo que o segundo valor representa quantas vezes esse valor foi verificado.

1. Para obter as medidas de variabilidade, no caso, variância e desvio-padrão, respectivamente:

> var (nome\_variável)

> sd (nome\_variável)

Ex. Calcule as medidas de variabilidade com a variável **Pessoas\_familia**

1. Comando que permite determinar o percentil, no caso o percentil 10

>quantile(nome\_variável,0.1)

Obs. Experimente usar o comando

>quantile(nome\_variável)

**Tabelas de frequência**

O comando “table” é utilizado para elaborarmos tabelas de frequências absolutas. Dependendo da variável a ser representada, podemos usar esse comando de diferentes formas:

1. Tabela simples para apresentação das frequências absolutas.

>table(nome\_variável)

Ex. Variável **Praticidade**

1. Tabela cruzada para apresentação das frequências absolutas.

>table(nome\_variável1,nome\_variável2)

Ex. Construir uma tabela cruzada apresentando as frequências absolutas das variáveis **Sexo** e **Divulgacao**.

1. Tabela cruzada para apresentação das frequências relativas.

>prop.table(table(nome\_variável1,nome\_variável2))

Ex. Construir uma tabela cruzada apresentando as frequências relativas das variáveis **Sexo** e **Divulgacao**.

**Gráficos**

1. Gráfico de colunas

As frequências podem ser visualizadas graficamente, usando gráficos de barras elementares, que se aplicam à descrição de qualquer variável qualitativa ou quantitativa discreta, vetor de dados ou tabelas.

No entanto, no caso de dados em banco de dados, quando não utilizamos outros mecanismos de atribuição, precisamos usar o comando table.

> barplot(table(nome\_variável))

Ex. Construir um gráfico de colunas para a variável **Sexo**.

Obs. É possível personalizar o gráfico, incluindo o título do eixo x (xlab), o título do eixoy (ylab), o título do gráfico (main), a cor da coluna (col) e cor da borda da coluna (border), lembrando que as cores, assim como os comandos devem ser expressas em inglês.

>barplot(table(nome\_variável),col=c("blue","red"),main="Título",xlab="Variável do eixo x", ylab="Informação que consta no eixo y", border='red')

Ex. Construir um gráfico de colunas para a variável **Pessoas\_familia**.

.

1. Setograma ou Gráfico de Pizza

>pie(table(nome\_variável),main="nome")

Ex. Construa um gráfico na forma de Setograma para a variável **Sabor**.

1. Para construção de um histograma, utilizado em geral quando temos variáveis quantitativas contínuas, devemos usar:

> hist(nome\_variável)

Ex. Construa um histograma com a variável **Renda\_h**.

Obs. Neste caso também é possível personalizar o gráfico, incluindo o título do eixo x (xlab), o título do eixoy (ylab), o título do gráfico (main), a cor da coluna (col) e cor da borda da coluna (border), lembrando que as cores, assim como os comandos devem ser expressas em inglês.

Obs. Para definir o número de intervalos no Histograma, usamos

>hist(nome\_variável, breaks = 5)

14) Para obter o boxplot para um conjunto de dados.

>boxplot(nome\_variável, names=c("A","B"))

Ex. Construir um boxplot da variável Idade.

Ex. Construir um boxplot das variáveis Peso\_(Kg) e Altura\_(m)

1. Gráfico Ramo e Folhas

>stem(nome\_variável) comando que permite obter um gráfico Ramo e Folhas.

Ou

>stem(nome\_variável,scale=1)

O 'scale=1', que é o padrão, separa os ramos das folhas a partir das casas decimais.

Caso padrão:

A ideia do ramo e folhas é separar um número (como 16,0) em duas partes. Assim, a primeira parte inteira (16) chamada de ramo e a segunda, a parte decimal (0) chamada de folha. O padrão do R é separar os números em duas partes (inteira e decimal) e agrupar os números em classes de tamanho 2. Por exemplo, o ramo 16 leva em conta os números 16 e 17.

**Obs**. Esse padrão vai se alterando, à medida que o conjunto de dados apresente diferentes casas decimais.

Assim, outras opções podem ser avaliadas:

a)>stem(nome\_variável,scale=0.5)

b) >stem(nome\_variável,scale=2)

Obs. Quando uma folha relacionada com certo ramo tem uma quantidade tão grande de valores que ele sintetiza essa quantidade usando a denominação +n, e invade a linha seguinte. Isso pode ser melhorado usando width.

c) >stem(nome\_variável,scale=0.5,width=120)

Ex. Construa um gráfico Ramo e Follhas com a variável **Idade**.