dsmath-practice

1 3 rawdatapractice.org

1.1 TIME TABLE

PRACTICE	MIN
Prerequisites	15
Loading packages	15
Looking at data	15
Factors vectors	15
Summary stats	15
Boxplots	10
Scatterplots	10
Bar charts	10
Customization	15
TOTAL	120

1.2 README

- Practice instructions for the course infrastructure
- Emacs + ESS + Org-mode and R must be installed
- **Make sure you sit at the same PC in every session**

- Upload the completed file as a class assignment

PRACTICE	MIN
Identify yourself	2
Find GitHub repos	2
Open the terminal	1
Open/close R from terminal	2
Emacs tutorial	2
Open/close R in Emacs	5
Run R in Org-mode file	15
Close Emacs/terminal	1
TOTAL	30

1.3 TODO Identify yourself

- Update the `#+AUTHOR:` information in the header
- Add this on a line to the header of this file : `#+STARTUP: overview hideblocks indent`
- With the cursor on theline, activate the header line with `C-c C-c`.
- Put your cursor on the headline of this section, and type `S <LEFT>` until you see `DONE` instead of `TODO` next to the title.
- Perform this last step each time you complete a section.

1.4 TODO Prerequisites

1. Check that R is installed on your machine. All of these are equivalent but lead to different interfaces:
   - open Emacs and open an R session with `M-x R`
   - open the CMD line terminal and enter the command `R`
   - open the CMD line terminal and enter the command `Rgui`

2. Check that you can execute R code blocks inside Emacs: execute the following code block named 1 by moving the cursor anywhere on the block - either on the metadata or on the line of code - and enter `C-c C-c`.

   The `result` of 1 is R `version` information. Because of the output size, it is automatically wrapped in an `example` block.

   ```
   version
   ```

3. Alternatively to executing the block in the Org-mode buffer, you can move the cursor on the block and enter `C-c '`. This will open the source code in a new buffer where you can execute it or edit it. This time, the output will appear in the `*R*` buffer instead of the Org-mode file.
4. If R is not installed, you need to install it. If you cannot execute R code blocks, you are probably missing the correct Emacs init file `/.emacs`: download the file from here. You might also miss the ESS (Emacs Speaks Statistics) package. Try `M-x load-library ESS RET`.
5. If you're in an R session now, exit by entering `q()`. R is case-sensitive, so this must be lower-case. When asked if you want to save the workspace, say **no**[1].

1.5 TODO Practice: data frames

1. Create and store this data frame as `dframe` in your R workspace.

   ```
     person sex funny
       Stan   M  High
   Francine   F   Med
      Steve   M   Low
      Roger   M  High
     Hayley   F   Med
      Klaus   M   Med
   ```

   The variables `person`, `sex`, and `funny` should be identical in nature to the variables in the `mydata` example:

   - `person` should be a **character** vector
   - `sex` should be a **factor** with levels `M` and `F`
   - `funny` should be a **factor** with levels `Low`, `Med`, and `High`.
2. Show that `sex` and `funny` are factors.

3. Show only the persons' names from the data frame.
4. Write one line of code that shows only the women whose funniness is at least `Med` or `High`. Use index operators or `subset`.

1.5.1 Solution

1. Create data frame.

```
dframe <- data.frame(
  person = c("Stan", "Francine", "Steve", "Roger", "Hayley", "Klaus"),
  sex = factor(c("M", "F", "M", "M", "F", "M")),
  funny = factor(c("High", "Med", "Low", "High", "Med", "Med")))
dframe
```

```
    person sex funny
1     Stan   M  High
2 Francine   F   Med
3    Steve   M   Low
4    Roger   M  High
5   Hayley   F   Med
6    Klaus   M   Med
```

1. Show the structure of data frame.

```
str(dframe)
```

```
'data.frame':   6 obs. of  3 variables:
 $ person: chr  "Stan" "Francine" "Steve" "Roger" ...
 $ sex   : Factor w/ 2 levels "F","M": 2 1 2 2 1 2
 $ funny : Factor w/ 3 levels "High","Low","Med": 1 3 2 1 3 3
```

2. Show only the names in the data frame.

```
dframe$person
```

```
[1] "Stan"     "Francine" "Steve"    "Roger"    "Hayley"   "Klaus"
```

3. Show only those women, whose funniness is at least `Med` or `High`.

```
## store relevant vectors
funny <- dframe$funny
sex <- dframe$sex

## extract elements with $ and [ ]
dframe$person[(funny=="Med" | funny=="High") & sex=="F"]

## extract elements with subset()
sub <- subset(x=dframe,(funny=="Med" | funny=="High") & sex=="F")
sub$person
```

```
[1] "Francine" "Hayley"
[1] "Francine" "Hayley"
```

```
class(dframe$funny)
```

```
[1] "factor"
```

1.6 TODO Practice: statistical variables

1. For each of the following, identify the **type of variable** described: numeric-continuous, numeric-discrete, categorical-nominal, categorical-ordinal

    1. The number of blemishes on the hood of a car coming off a production line
    2. A survey question that asks the participant to select from: "Strongly agree", "Agree", "Neutral", "Disagree", and "Strongly disagree"
    3. The noise level (in decibels) at a concert
    4. The noise level out of three possible choices: high, medium, low
    5. A choice of primary color
    6. The distance between a cat and a mouse

VARIABLE	TYPE
1	
2	
3	
4	
5	
6	

2. For each of the following, identify whether the quantity discussed is a population **parameter** or a sample **statistic**. If the latter, also identify what the corresponding population parameter is.

    1. The percentage of 50 New Zealanders who own a gaming console.
    2. The average number of times per day a vending machine is used in a year
    3. The proportion of domestic cats in the United States that wear a collar
    4. The average number of times per day a vending machine is used in a year
    5. The average number of times per day a vending machine is used in a year, based on data collected on three distinct days in that year.

CHAR	TYPE	CORRESPONDING PARAMETER
1		
2		
3		
4		
5		

1.6.1 Solution

1. For each of the following, identify the **type of variable** described: numeric-continuous, numeric-discrete, categorical-nominal, categorical-ordinal

    1. The number of blemishes on the hood of a car coming off a production line - **numeric-discrete** in [0,N]
    2. A survey question that asks the participant to select from: "Strongly agree", "Agree", "Neutral", "Disagree", and "Strongly disagree" - **categorical-ordinal**
    3. The noise level (in decibels) at a concert - **numeric-continuous**
    4. The noise level out of three possible choices: high, medium, low - **categorical-orginal**
    5. A choice of primary color - **cateogircal-nominal**
    6. The distance between a cat and a mouse - **numeric-continuous**

VAR	TYPE
1	numeric-discrete
2	categorical-ordinal
3	numeric-continuous
4	categorical-ordinal
5	categorical-nominal
6	numeric-continuous

2. For each of the following, identify whether the quantity discussed is a population **parameter** or a sample **statistic**. If the latter, also identify what the corresponding population parameter is.

    1. The percentage of 50 New Zealanders who own a gaming console.
    2. The average number of times per day a vending machine is used in a year
    3. The proportion of domestic cats in the United States that wear a collar
    4. The average number of times per day a vending machine is used in a year
    5. The average number of times per day a vending machine is used in a year, based on data collected on three distinct days in that year.

CHAR	TYPE	CORRESPONDING PARAMETER
1	statistic	percentage who own console
2	statistic	average no. of uses
3	parameter	
4	parameter	
5	statistic	average no. of uses

1.7 TODO Test questions

You now should be able to answer these test questions. You can find short answers in the footnote[2]:

1. What is a data frame? How can you create one?
2. What is the difference between a vector and a factor?
3. How can you extract a range of rows or columns from a data frame?
4. What is a population, a parameter, a sample and a statistic?

5. What is the purpose of statistics?

1.7.1 Solution

1. What is a data frame? How can you create one?
    - Table of column vectors of same length
    - `data.frame` function
2. What is the difference between a vector and a factor?
    - a factor has nominal or ordinal levels, finite, discrete values
3. How can you extract a range of rows or columns from a data frame?
    - Index operators $ and [ ] or subset function
4. What is a population, a parameter, a sample and a statistic?
    - population: what we're after, parameter: population characteristic, sample: what we can collect, statistic: sample characteristic
5. What is the purpose of statistics?
    - Techniques to draw conclusions on populations from samples

Footnotes:

[1] If you say yes, R will save a copy of all your commands in that session in a file `.Rhistory`, and it will save all data in a file `.RData` to recreate your work space the way you left it.

[2] 1) an R object that consists of vectors of the same length, created with the `data.frame` function. 2) factors are vectors of categorical values with ordered or unordered levels. 3) Using the accessor operator `$` or the index operator `[]`, where the accessor requires a named, non-atomic, vector. 4) A population parameter is a characteristic of interest in something in the world; a sample statistic is an estimate of the population parameter based on a sample, a subset drawn at random from the population. 5) Statistics allow to infer population characteristics from sample characteristics.

Created: 2022-09-22 Thu 15:11