STA501 Week #3 Assignment

Due: 11:30 PM, Sunday, 2/14/2021

2/9/2021

Contents

1 Introduction and Data Source

This week's assignment focuses on the descriptive statistics using R. The **Diabetes** data set to be used in this assignment is taken from Vanderbilt's Biostatistics Datasets.

The following is the description from the web page:

These data are courtesy of Dr. John Schorling, Department of Medicine, University of Virginia School of Medicine. The data consist of 19 variables on 403 subjects from 1046 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. According to Dr. John Hong, Diabetes Mellitus Type II (adult-onset diabetes) is associated most strongly with obesity. The waist/hip ratio may be a predictor of diabetes and heart disease. DM II is also associated with hypertension - they may both be part of "Syndrome X". The 403 subjects were the ones who were actually screened for diabetes. Glycosolated hemoglobin > 7.0 is usually taken as a positive diagnosis of diabetes. For more information about this study see

Willems JP, Saunders JT, DE Hunt, JB Schorling: Prevalence of coronary heart disease risk factors among rural blacks: A community-based study. *Southern Medical Journal* 90:814-820; 1997

Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL: A trial of church-based smoking cessation interventions for rural African Americans. *Preventive Medicine* 26:92-101; 1997.

```r
diabetes <- read.csv("https://hbiostat.org/data/repo/diabetes.csv", header = TRUE)
kable(head(diabetes))
```

id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
1000	203	82	56	3.6	4.31	Buckingham	46	female	62	121	medium	118	59	NA	NA	29	38	720
1001	165	97	24	6.9	4.44	Buckingham	29	female	64	218	large	112	68	NA	NA	46	48	360
1002	228	92	37	6.2	4.64	Buckingham	58	female	61	256	large	190	92	185	92	49	57	180
1003	78	93	12	6.5	4.63	Buckingham	67	male	67	119	large	110	50	NA	NA	33	38	480
1005	249	90	28	8.9	7.72	Buckingham	64	male	68	183	medium	138	80	NA	NA	44	41	300
1008	248	94	69	3.6	4.81	Buckingham	34	male	71	190	large	132	86	NA	NA	36	42	195

We can see from the first 6 observations that there are 15 numerical variables and 3 categorical variables. Variable **bp.2s** and **bp.2d** have missing values. To complete this week's assignment, you need to choose one numerical variable and one categorical variable with NO missing values.

The following code shows how to extract variables from the data frame. I will use the two variables with missing values as an example. You can modify the code to extract your variables for the assignment.

```
bp.2s <- diabetes$bp.2s
bp.2d <- diabetes$bp.2d
```

2 Assignments for This Week: Due: 11:30 PM, Sunday, 2/14/2021

This assignment focuses on descriptive statistics on categorical and numerical data. Please prepare an R Markdown document to complete the assignment. You are expected to submit the RMarkdown document and one of the three converted documents: HTML, PDF, and Word.

2.1 Summarizing Categorical Data

Use the categorical variable you selected to perform the following analysis

1. **Construct a relative frequency table**. write a few sentences to describe the distribution of the variable. Note that you are encouraged to construct a frequency table with all four types of frequencies as I did in the class note.
2. **Construct a pie-chart** to represent the distribution of the categorical variable.

2.2 Summarizing Numerical Data

Using the numerical variable you chose from the diabetes data to answer the following questions.

3. **Construct a relative frequency table of the numerical variable** with 10 categories. In other words, the frequency table should have 10 rows. You are encouraged to include all 4 frequencies in the table. Please provide a brief description of the relative frequencies.
4. **Construct a histogram of the numerical variable** with 10 vertical bars. In other words, the histogram is a geometric representation of the frequency table. Explain the distribution of the variable. Is it skewed to the left or the right?
5. **Construct a box-plot** and explain it. That is, can you tell whether the distribution is skewed to the right or the left?