

descriptive statistics

INTERMEDIATE ANALYTICS ASSIGNMENT 1



Descriptive statistics is a method used to describe, show and summarize data such that meaningful solutions can be derived from it by detecting patterns or trends in the data. It is merely a way to describe the data and does not necessarily provide any ways to draw concrete conclusions from the data because it only tells us what we need to know about that specific dataset. It is used across various industries to determine the behavior of specific data such as consumer behavior. There are two types of namely measure of central tendency which allows you to find the average, mean and mode of a population and measure of dispersion which explores the data range, skewness and standard deviation.

May 27, 2018

Victoria plange

ALY 6015 SEC 03

PROFESSOR JUSTIN RODGERS

Question 1

* Download the attached dataset “assignment1.csv, set your working directly to the location of the file, and import the dataset using R code. Using descriptive statistics, visualize this data numerically and graphically. In particular, please provide summary statistics (mean, median, standard deviation, min, max, Q1, and Q3) for variable “Sepal.Width”. Also, please create a histogram for “Sepal.Width”. You can use the following resource as a guide: section 2.5 (p. 15) and 3.5 (p.25) from “Using R for Data Analysis and Graphics.”

Solution

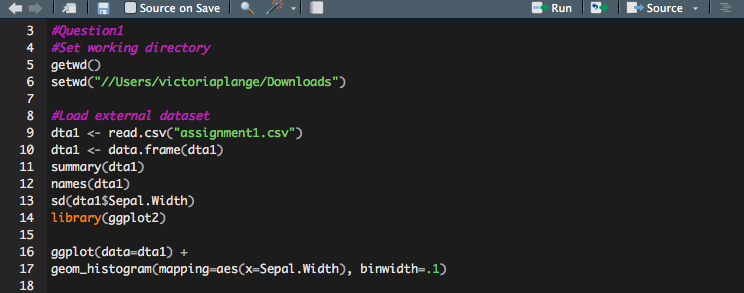


Fig.1.1

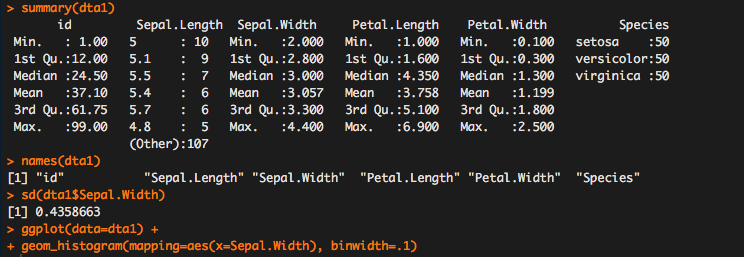


Fig.1.2

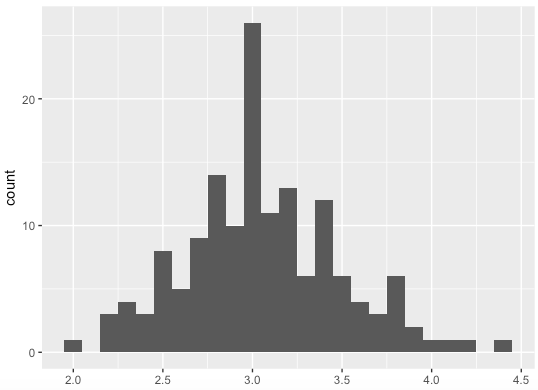


Fig.1.3

Fig.1.1 provides a summary of our data set for Iris flower species- setosa, versicolor and virginica. The summary function in R displays the minimum, maximum, first quartile, third quartile, median and mean of the column within each data set. For Sepal.Width, we found the mean and median to be close together, with the mean being slightly larger than the median (3.057) allowing us to anticipate a skewed right distribution. The standard deviation measures how spread out our values are from the mean and if they are further from the mean, it shows a higher variation in the data set. Within Sepal.Width, we determined a stdev of 0.4 telling us that are data points lie closer to the mean. The data was then visualized into a histogram via ggplot package in R. A histogram represents an accurate distribution and depiction of our numerical data. Fig 1.3 shows the histogram for Sepal.Width and it is a right skewed distribution. We see that we have about 26 species with a sepal width of 3 and between 8 and roughly 18 species with a width between 2.5 and 3.7. The plant width size determines the leaf size and tells us how the leaves ability to absorb light which is an essential factor for their growth. We can ultimately determine from the data that the plants with a sepal width of 3.0 and above would have more energy to process food.

Question 2&3

* Using  [height\_weight\_byGender.csv](https://northeastern.blackboard.com/bbcswebdav/pid-18179710-dt-content-rid-41422766_1/xid-41422766_1) data file, plot a histogram of men’s and women’s height.
* Then, again using the height\_weight\_byGender.csv file, plot a scatterplot of height by weight for either men or women using the plot() R function.

Solution



Fig.1.4

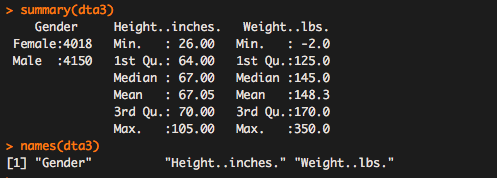


Fig.1.5

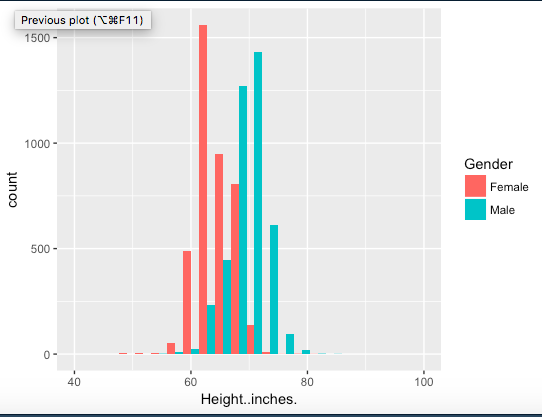


Fig.1.6

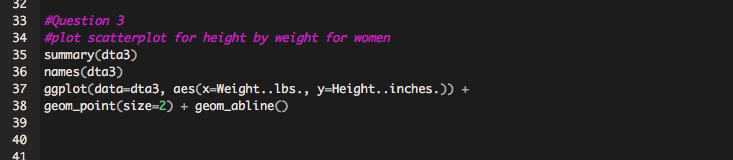


Fig.1.7

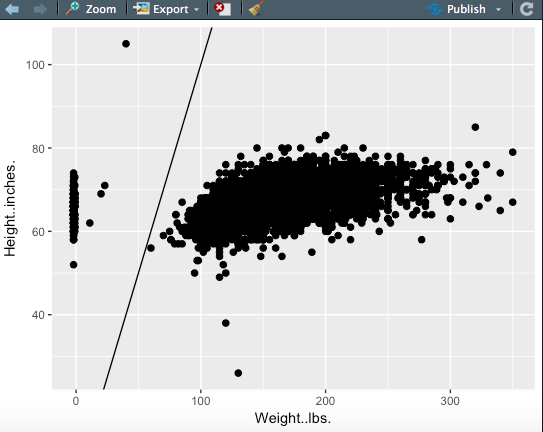


Fig.1.8

Fig.1.5 shows a summary of our dataset, height\_weight\_bygender which tells us different male and female heights and their corresponding weights. The name function gives us a list of the column names within the data set namely Gender, Height..inches, and Weight.. lbs. Using ggplot package in R, we plot a histogram to visualize the summary of height in inches for both males and females shown in Fig.1.6. We can tell from the image that males have a higher height in inches compared to females. When this same data is translated into a scatter plot, as seen in Fig.1.8 we determine a correlation and a positive linear relationship between height and weight, thus as the height increases the weight also increase. However, this is not so linear based on the fact that the line does not go through the points. We also detected the existence of outliers, which has had an effect on the median of the data and thus the skewness. The outliers also determine that as the height increases rapidly, the weight also increases but at a slower rate.

Question 4

Descriptive statistics is all about describing a sample of data (or a population). Above, you utilized descriptive statistics to describe the height and weight of individuals within the “height\_weight” dataset. What is an example of a statistical inference question we may ask next?

Solution

* Inferential statistics is when we have a limited amount of data and can use that to make assumptions about the entire population. Given the dataset on weight and height of males and females within a given location, we can determine that an increase in height and a slow increase in weight can tell us that there is an underlying factor causing the slow growth in weight. Perhaps a malnourished diet, where certain nutrients for weight gain and strength are absent, could affect the weight gain in both males and females.

References

* [Iris Flower Species data set]. (n.d.). Unpublished raw data.
* [Height\_Weight\_ByGender]. (n.d.). Unpublished raw data.