Experiment1-Descriptive Statistics

Descriptive statistics is a set of math used to summarize data. Descriptive statistics can be distribution, central tendency, and dispersion of data.The distribution can be a normal distribution or binomial distribution. The central tendency can be mean, median, and mode. The dispersion or spreadness can be the range, interquartile range, variance, and standard deviation.In this session, you will import a CSV file, Excel file and you will perform basic data processing. I will explain descriptive statistics, central tendency measurements, dispersion measurements. You will look into how R programming can be used to calculate all these values.

What Is Descriptive Statistics?

Descriptive statistics summarizes the data and usually focuses on the distribution, the central tendency, and dispersion of the data. The distributions can be normal distribution, binomial distribution, and other distributions like Bernoulli distribution. Binomial distribution and normal distribution are the more popular and important distributions, especially normal distribution. When exploring data and many statistical tests, you will usually look for the normality of the data, which is how normal the data is or how likely it is that the data is normally distributed. The Central Limit Theorem states that the mean of a sample or subset of a distribution will be equal to the normal distribution mean when the sample size increases, regardless whether the sample is from a normal distribution. The central tendency, not the central limit theorem, is used to describe the data with respect to the center of the data. Central tendency can be the mean, median, and mode of the data. The dispersion describes the spread of the data, and dispersion can be the variance, standard deviation, and interquantile range. Descriptive statistics summarizes the data set, lets us have a feel and understanding of the data and variables, and allows us to decide or determine whether we should use inferential statistics to identify the relationship between data sets or use regression analysis to identify the relationships between variables.

Reading Data Files

R programming allow you to import a data set, which can be comma-separated values (CSV) file, Excel file, tab-separated file, JSON file, or others. Reading data into the R console or R is important, since you must have some data before you can do statistical computing and understand the data. Before you look into importing data into the R console, you must determine your workplace or work directory first. You should always set the current workspace directory to tell R the location of your current project folder. This allows for easier references to data files and scripts.

To print the current work directory, you use the getwd() function:

```
# get the current workspace location
print(getwd());
> print(getwd());
[1] "C:/Users/gohmi/Documents"
```

```
#set the current workspace location
```

```
setwd("D:/R"); #input your own file directory, for here
we use "D:/R"
> setwd("D:/R");
```

To get the new work directory location, you can use the getwd() function:

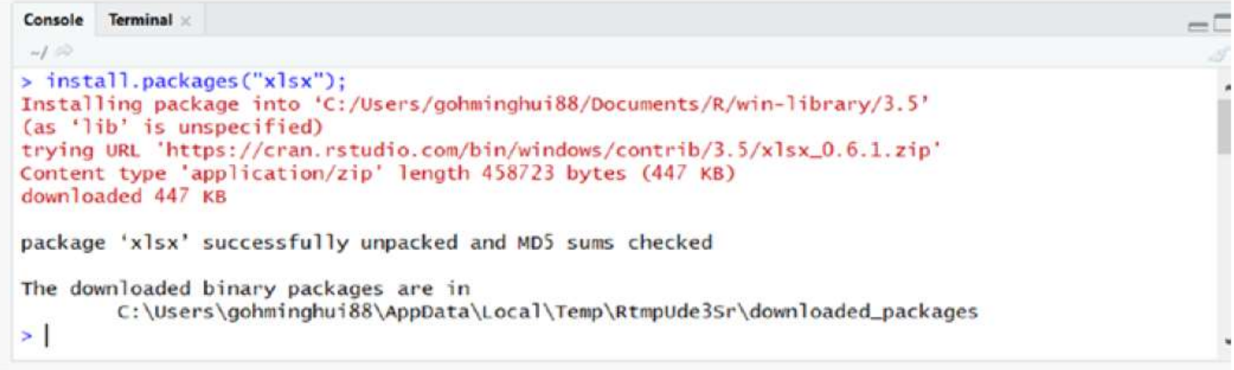
```
#get the new workspace
print(getwd());
> print(getwd());
[1] "D:/R"
```

You can put the data.csv data set into D:/R folder.

Reading an Excel File

The data set can also be in the Excel format or .xlsx format. To read an Excel file, you need to use the xlsx package. The xlsx package requires a Java runtime, so you must install it on your computer. To install the
xlsx package, go to the R console and type the following, also shown in Figure

```
> install.packages("xlsx");
```



```
~/
> install.packages("xlsx");
Installing package into ‘C:/Users/gohminghui88/Documents/R/win-library/3.5’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/xlsx_0.6.1.zip'
Content type 'application/zip' length 458723 bytes (447 KB)
downloaded 447 KB

package ‘xlsx’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\gohminghui88\AppData\Local\Temp\RtmpUde3Sr\downloaded_packages
> |
```

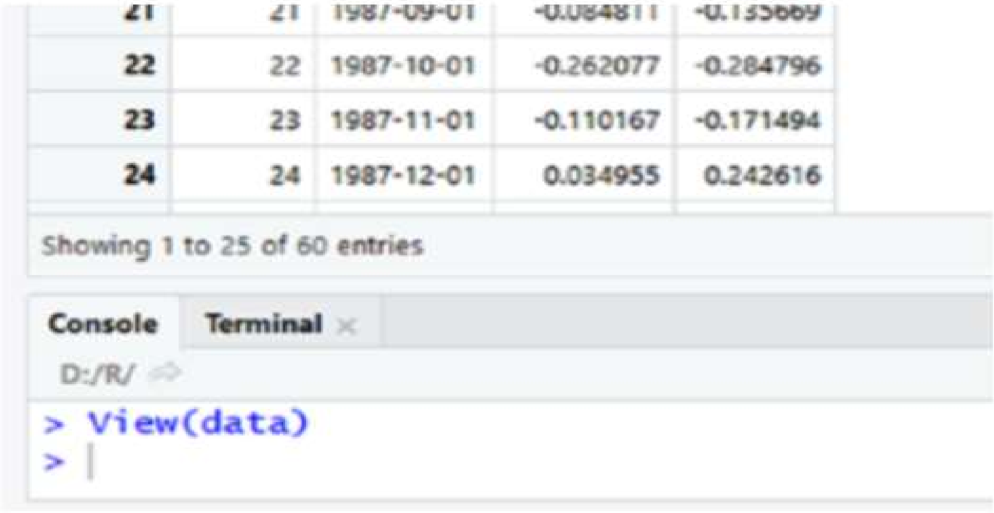
To use the xlsx package, use the require() function:

```
> require("xlsx");
Loading required package: xlsx
```

To read the Excel file, you can use the read.xlsx() function:

```
> data <- read.xlsx(file="data.xlsx", 1);
```

file is the location of the Excel file. 1 refers to sheet number 1.To view the data variable, you can use the View() function or click the data variable in the Environment portion of RStudio, as shown in Figure.



21	21	1987-09-01	-0.084811	-0.135669
22	22	1987-10-01	-0.262077	-0.284796
23	23	1987-11-01	-0.110167	-0.171494
24	24	1987-12-01	0.034955	0.242616

Showing 1 to 25 of 60 entries

```
> View(data)
>
```

To look for the documentation of read.xlsx(), you can use the following code.

```
> help(read.xlsx);
```

The data variable is of the data frame data type:

```
> class(data);
[1] "data.frame"
```

Writing an Excel File

To write a Excel file, you can use the write.xlsx() function:

```
> write.xlsx(data, file="data2.xlsx", sheetName="sheet1", col.names=TRUE,
row.names=FALSE);
```

data is the variable of data frame type to export to Excel file, file is the file location or path, sheetName is the sheet name, and col.names and row.names are logical values to state whether to export with column names or row names. To view the documentation of the write.xlsx() function or any R function, you can use the help() function.

Basic Data Processing

After importing the data, you may need to do some simple data processing like selecting data, sorting data, filtering data, getting unique values, and removing missing values.

```
data=read.csv("C:/Users/dkalp/OneDrive/Desktop/spreadsheet.csv")
```

Mode, Median, Mean

Mean, median, and mode are the most common measures for central tendency. Central tendency is a measure that best summarizes the data and is a measure that is related to the center of the data set.

Mode

Mode is a value in data that has the highest frequency and is useful when the differences are non-numeric and seldom occur.

To get the mode in R, you start with data:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8); #To get mode in a vector, you create a frequency table:
> y <- table(A);
> y;
A
1 2 3 4 5 6 7 8
1 1 1 1 3 1 1 1
```

You want to get the highest frequency, so you use the following to get the mode:

```
> names(y)[which(y==max(y))];
[1] "5"
```

Median

The median is the middle or midpoint of the data and is also the 50 percentile of the data. The median is affected by the outliers and skewness of the data. The median can be a better measurement for centrality than the mean if the data is skewed. The mean is the average, which is liable to be influenced by outliers, so median is a better measure when the data is skewed.

In R, to get the median, you use the median() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> median(A);
[1] 5
```

Mean

The mean is the average of the data. It is the sum of all data divided by the number of data points. The mean works best if the data is distributed in a normal distribution or distributed evenly. The mean represents the expected value if the distribution is random.

In R, to get the mean, you can use the mean() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> mean(A);
[1] 4.6
```

Handle NA Values with mean Function

A typical problem occurs when the data contains NAs. Let’s modify our example vector to simulate such a situation:

```
>B=c(A,NA)
>B
[1]  1  2  3  4  5  5  5  6  7  8 NA
```

Our new example vector looks exactly the same as the first example vector, but this time with an NA value at the end. Let’s see what happens when we apply the mean function as before:

```
>mean(B)
> [1] NA
```

The RStudio console returns NA – not as we wanted. Fortunately, the mean function comes with the na.rm (i.e. NA remove) option, which can be used to ignore NA values. Let’s do this in practice:

```
>mean(B,na.rm=TRUE)
>[1] 4.6
```

As you can see, we get the same mean output as before.
Note:The na.rm option can also be used to ignore `NaN` or `NULL` values.

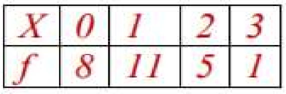
Problem1:Twenty students , graduates and undergraduates, were enrolled in a statistics course. Their ages were
18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36.
a) Find Mean and Median of all students
b) Find median age of all students under 25 years.
c) Find modal age of all student

R code:- >

```
x=c(18,19,19,19,19,20,20,20,20,20,21,21,21,21,22,23,24,27,30,36)
 > mean(x) #mean
[1] 22
> median(x) #median
[1] 20.5
> y=x[x<25]
 >median(y)
[1] 20
> xr=table(x) #mode
> mode=which(xr==max(xr))
> mode
20
3
```

Measures of central tendency for frequency table:-

Problem 2 : A survey of 25 faculty members is taken in a college to study their vocational mobility.They were asked the question “In addition to your present position ,at how many educational insistutes have served on the faculty?.Following is the frequency distribution of their responses .



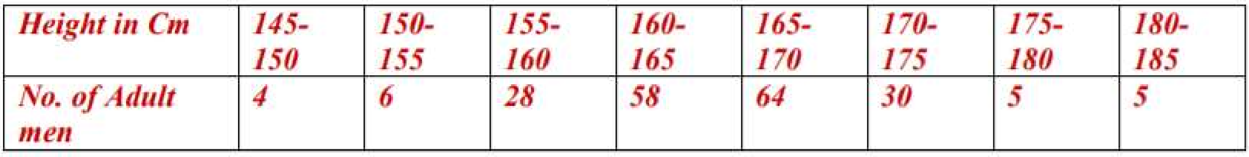
X	*0*	*1*	*2*	*3*
f	*8*	*11*	*5*	*1*

Find mean and median of the distribution
R code:

```
> x=c(0,1,2,3)
> f=c(8,11,5,1)
> y=rep(x,f)
> mean=(sum(y))/(length(y)) #mean
 > mean
```

```
[1] 0.96
> median(y) #median
[1] 1
```

Problem 3 : Compute mean ,median , 1st Quartile, 3rd Quartile and mode of for the following frequency Distribution:



Height in Cm	145-150	150-155	155-160	160-165	165-170	170-175	175-180	180-185
No. of Adult men	4	6	28	58	64	30	5	5

```
> x=seq(147.5,182.5,5)
> x
[1] 147.5 152.5 157.5 162.5 167.5 172.5 177.5 182.5
> f=c(4,6,28,58,64,30,5,5)
> mean=sum(x*f)/sum(f)
> mean
[1] 165.175

For Median:

> c=cumsum(f)
> cl=cumsum(f)
> cl
[1]   4  10  38  96 160 190 195 200
> N=sum(f)
> N
[1] 200
> ml=min(which(cl>N/2))
> ml
[1] 5
> h=5
> h
[1] 5
> fm=f[ml]
> fm
[1] 64
> cf=cl[ml-1]
> cf
[1] 96
> l=x[ml]-h/2
> l
[1] 165
> median=l+(((N/2)-cf)/fm)*h #median
> median
[1] 165.3125
```

To find Quartile 1:

```
> Q1=min(which(cl>N/4))
> Q1
[1] 4
> fq1=f[Q1]
> fq1
[1] 58
> cf1=cl[Q1-1]
> cf1
[1] 38
> l=x[Q1]-h/2
> l
[1] 160
> quartile1=l+(((N/4)-cf1)/fq1)*h
> quartile1
[1] 161.0345
```

To find Quartile 3:

```
> Q3=min(which(cl>3*N/4))
> Q3
[1] 5
> fq3=f[Q3]
> fq3
[1] 64
> cf2=cl[Q3-1]
> cf2
[1] 96
> l=x[Q3]-h/2
> l
[1] 165
> quartile3=l+(((3*N/4)-cf2)/fq3)*h
> quartile3
[1] 169.2188
```

Mode:

```
> m=which(f==max(f))
> m
[1] 5
> f0=f[m]
> f0
[1] 64
> f1=f[m-1]
> f1
[1] 58
> f2=f[m+1]
> f2
[1] 30
> l=x[m]-h/2
> l
[1] 165
> mode=l+((f0-f1)/(2*f0-f1-f2))*h
> mode
[1] 165.75
```

Range,Interquartile Range, Variance, Standard Deviation

Measures of variability are the measures of the spread of the data. Measures of variability can be range, interquartile range, variance, standard deviation, and more.

Range

The range is the difference between the largest and smallest points in the data.
To find the range in R, you use the *range()* function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> range(A);
[1] 1 8
```

To get the difference between the max and the min, you can use

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> res <- range(A);
> diff(res);
[1] 7
```

You can use the min() and max() functions to find the range also:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> min(A);
[1] 1
> max(A);
[1] 8
> max(A) - min(A);
[1] 7
```

To get the range for a data set:

```
> diff(res);
[1] 10.65222
```

Interquartile Range

The interquartile range is the measure of the difference between the 75 percentile or third quartile and the 25 percentile or first quartile.
To get the interquartile range, you can use the IQR() function:

```
> A <- c(1, 2, 3, 4, 5, 5, 5, 6, 7, 8);
> IQR(A);
[1] 2.5
```

You can get the quartiles by using the quantile() function:

```
> quantile(A);
0% 25% 50% 75% 100%
1.00 3.25 5.00 5.75 8.00
```

You can get the 25 and 75 percentiles:

```
> quantile(A, 0.25);
25%
3.25


> quantile(A, 0.75);
75%
5.75
```

The IQR() and quantile() functions can have NA values removed using na.rm = TRUE.

Range measures the maximum and minimum data value , and the interquartile range measures where the majority value is.

Example:
An entomologist studying morphological variation in species of mosquito recorded the following data on body length: 1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3.Compute all the measures of disersion.

```
> x=c(1.2,1.4,1.3,1.6,1.0,1.5,1.7,1.1,1.2,1.3)
> x
 [1] 1.2 1.4 1.3 1.6 1.0 1.5 1.7 1.1 1.2 1.3
> res=range(x)
> res
[1] 1.0 1.7
> diff(res)
[1] 0.7
> var(x)   # Variance
[1] 0.049
> sd(x)    # standard deviation
[1] 0.2213594
> quantile(x)
 0%    25%    50%    75%   100%
1.000 1.200 1.300 1.475 1.700

First Quartile  is 1.2
Second Quartile is 1.3
Third quartile is 1.475

> IQR(x)   # Inter quartile range
[1] 0.275
```

Mean deviation about Mean, Median and Mode:

```
> y=abs(x-mean(x))
> M1=sum(y)/length(y)  # mean deviation about Mean
> M1
[1] 0.176
> z=abs(x-median(x))
> M2=sum(z)/length(z)  # Mean deviation about median
> M2
[1] 0.17
Mean deviation about Mode # in this Problem ,it is a bi-model series (Mode is not
possible)
```

References

1. Biological data analysis, Tartu 2006/2007 (Tech.). (n.d.). Retrieved September 1, 2018, from www-1.ms.ut.ee/BDA/BDA4.pdf.

2. Calculate Standard Deviation. (n.d.). Retrieved from https://explorable.com/calculate-standard-deviation.

3. Descriptive Statistics. (n.d.). Retrieved from http://webspace.ship.edu/cgboer/descstats.html.

4.Descriptive statistics. (2018, August 22). Retrieved from https://en.wikipedia.org/wiki/Descriptive_statistics.

5.Donges, N. (2018, February 14). Intro to Descriptive Statistics – Towards Data Science. Retrieved from https://towardsdatascience.com/intro-to-descriptive-statistics-252e9c464ac9.

6. How to Make a Histogram with Basic R. (2017, May 04). Retrieved from www.r-bloggers.com/how-to-make-a-histogram-with-basic-r/.