Ch 1 : Descriptive Statistics

Contents

June 9, 2017

1

Preliminaries

[ToC]

1.1 Course Web Page

[ToC]

- Course web page: `http://gozips.uakron.edu/~nmimoto/461`

- Or search "Mimoto" in UA web site.

- We also uses UA Springborad for HWs and grade postings.

1.2 Statistical Software - R

- We will use Statistical Software called "R" in our class and some of our HW.

- It is a free software, can be installed in PC or MAc

- Similar and as powerfull as Matlab

- Official Website: `https://www.r-project.org`

  Click CRAN $\rightarrow$ (pick mirror close to OH) $\rightarrow$ Download R for Windows (Mac)

1.3 Importing Data to R

Method 1: Download dataset directly to R from Course Web Page.

- Open R, and run the following command in R

```
D1  <- read.csv("http://gozips.uakron.edu/~nmimoto/pages/datasets/pi.csv")
Pi  <- as.numeric(D1[,1])  #- turn it into numbers

head(Pi)    #- see first few lines

plot(Pi)    #- Scatter plot

hist(Pi, (0:10)-.5)    #- Histogram

View(Pi)          #- Opens spreadsheet view

Pi <- edit(Pi)   #- data editor
```

Importing Data to R - Method 2

Download dataset to your PC, then load it to R.

- Download `.csv` file from course web site to your PC.
- Make sure the file is stored in your working directory (use getwd() command)
- (If not, go File → Change dir...)
- Open up R, and type

```
D1 <- read.csv("pi.csv")
Pi  <- as.numeric(D1[,1])
```

- Or you can give path to the data file without changing your working directory.

```
D1 <- read.csv("C:\Users\mimoto\Documents\pi.csv")
Pi  <- as.numeric(D1[,1])
```

When it is already inside R

R has many dataset already loaded inside. To load built-in data, use data() command.

```
data(quakes)    #- load the built-in data "quakes"

quakes

help(quakes)    #- opens description of the data
```

2

Descriptive Statistics

[ToC]

2.1 Types of data

- Univariate: Records one variable.
- Bivariate (Multivariate): Records more than one variable.

Load Datasets

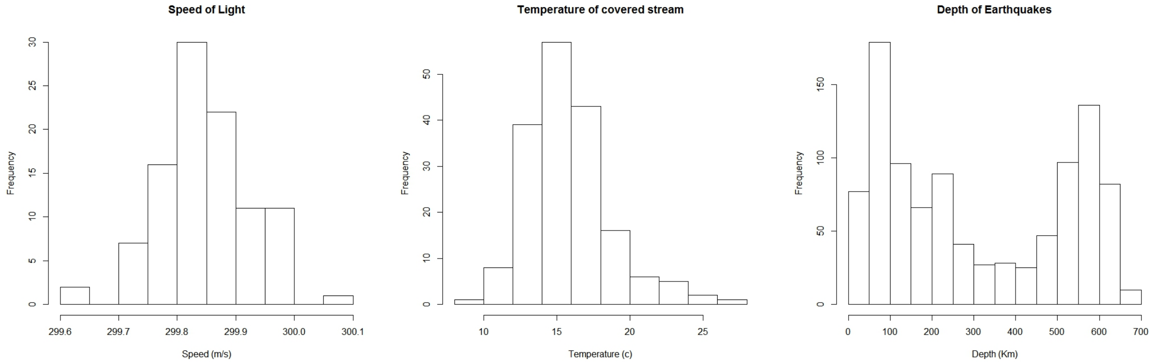
```
D1  <- read.csv("http://gozips.uakron.edu/~nmimoto/pages/datasets/light.csv", header=T)
Light <- as.numeric(D1[,1])

D2  <- read.csv("http://gozips.uakron.edu/~nmimoto/pages/datasets/stream.csv", header=T, skip=1)
Temp <- as.numeric(D2[,2])

Depth <- as.numeric( quakes[,"depth"] )
```

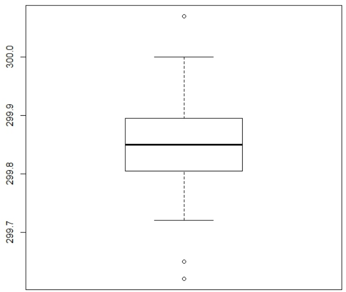
2.2 Histogram and its shape

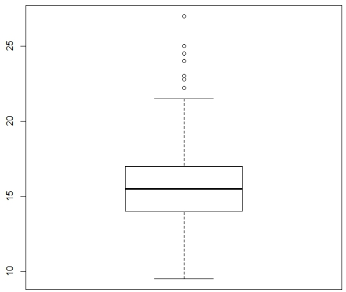
- Unimodal, Right skewed, Bimodal (Multimodal).

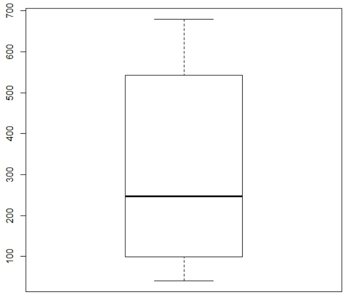



```
hist(Light, main="Speed of Light", xlab="Speed (m/s)")
hist(Temp, main="Temperature of covered stream", xlab="Temperature (c)")
hist(as.numeric(  quakes[,"depth"]  ), main="Depth of Earthquakes", xlab="Depth (Km)")
hist(as.numeric(  quakes[,"mag"]  ), main="Magnitude of Earthquakes", xlab="Magnitude")
```

2.3 Boxplots

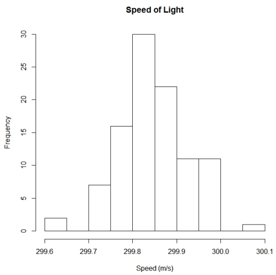


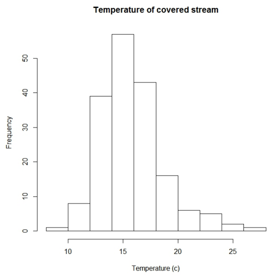


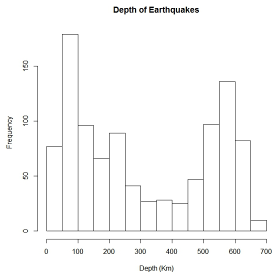


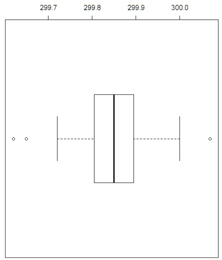

```
boxplot(Light)
boxplot(Temp)
boxplot(Depth)
```

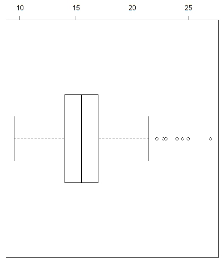
Compare Histograms and Boxplots

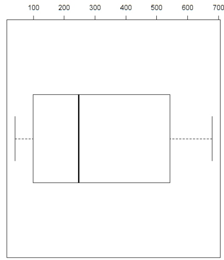












2.4 Five number summary of data

Min	Q1	Median	Q3	Max
	1st quartile	2nd quartile	3rd quartile	
	25th percentile	50th percentile	75th percentile	100th percentile

Boxplot is drawn using these five numbers. (get quartiles by INCLUDING median)

Sample median is

$$\widetilde{x} = \begin{cases} \frac{n+1}{2}\text{th ordred observations} & \text{if n is odd} \\ \text{average of } \left(\frac{n}{2}\right) \text{ th and } \left(\frac{n}{2}+1\right) \text{ th ordered observations} & \text{if n is even} \end{cases}$$

IQR: InterQuartile Range is (Q3 - Q1).

Example:

- Suppose our data look like this:

$$1, 2, 3, 4, \underbrace{5}_{\text{median}}, 6, 7, 8, 9$$

  Then Q1 is the median of the lower half (including the median), which is 3.

  Similarly, Q3 is 7.

- Suppose our data look like this:

$$1, 2, 3, 4, \underbrace{5, \quad 6}_{\text{median=5.5}}, 7, 8, 9, 10$$

  Q1 is still 3, because the median of the lower half (1 through 5). Q3 is now 8.

```
summary(Depth)
boxplot(Depth)
```

2.5 Boxplot and Outliers

You can use 5 number summary to draw a box-plot.

Outlier Observations further than 1.5 box width away from the closest fourth is an outlier. If it is more than 3 box width away from the nearest fourth, it's called extreme outlier. Otherwise it is called an mild outlier.

2.6 Mean, Variance, and Std Dev

Let $X_1, X_2, X_3, \ldots, X_n$, be a random sample of size n. Then,

- **Sample mean** is

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

- **Sample vaiance** is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

  Notice that we are dividing by $n - 1$ instead of n.

- **Sample Standard Deviation** is defined as

$$s = \sqrt{s^2}.$$

- Mean is the "center" of the data. If you draw a histogram, it is where you can balance the histroram.
- Median is also "center" of the data. If you draw a histogram, area of the histogram is same for left of the median and right of the median.
- Variance and SD is the "width" of the data.

```
mean(Light)   #- mean

var(Light)    #- variance

sd(Light)     #- standard deviation
```

2.7 Calculate variance by hand

Suppose the dataset is

$$3.2,\ 4.5,\ 5.5,\ 6.1,\ 10.7$$

To calculate the variance by hand, we must first calculate the mean, $\bar{X} = 6$. Then we have to make a table like below:

i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	3.2	-2.8	7.84
2	4.5	-1.5	2.25
3	5.5	-0.5	0.25
4	6.1	0.1	0.01
5	10.7	4.7	22.09

Adding up the last column and dividing by (5-1)=4 gives variance as 8.11.

SD is $\sqrt{8.11} = 2.848$.

2.8 Mean and Median for skewed data

- On right skewed data, mean is **larger** than median.

2.9 Mean vs Median on sensitivity

- Suppose our data look like this:

$$1, 2, 3, 4, 5$$

  Then mean is 3, median is also 3.

- Suppose our data look like this:

$$1, 2, 3, 4, 50$$

  Now mean is 12, but median is still 3.

- Mean can be changed a lot by one number. Median does not. Median is more robust than mean. In other words, mean is sensitive.