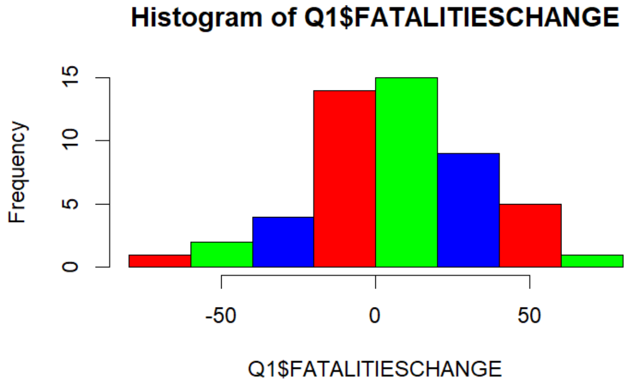
STAT 40001/STAT 50001 Statistical Computing Fall 2024

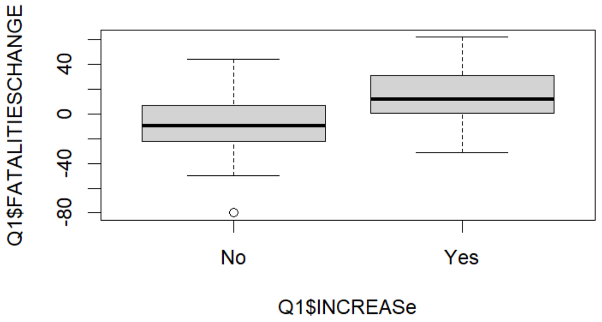
Lab-7

1) The National Highway System Designation Act was signed into law in 1995. It abolished the federal mandate of 55 mph speed limits. The data ***speed** (provided in the Brightspace)* shows percentage changes in interstate highway traffic fatalities from 1995 to 1996 Important note: Please note that it is a tab-delimited file.

a) Print first 5 lines of the data
b) Draw the histogram of the percentage changes in interstate highway traffic fatalities from 1995 to 1996
c) Compare the speed limit and traffic fatalities by displaying a side by side boxplots

```
> Q1 <- read.table("C:\\Users\\PNW_checkout\\Downloads\\vaishak\\PNW_COURSE-WORK\\FALL24\\STATIST
ICAL COMPUTING\\Assignment\\Assignment 7\\speed.txt", sep = "\t", header = T)
> head(Q1, 5)
              STATE INCREASe FATALITIESCHANGE
1            Alaska       No            -29.0
2       Connecticut       No             -4.4
3 Dist. of Columbia       No            -80.0
4            Hawaii       No            -25.0
5           Indiana       No            -13.2
> hist(Q1$FATALITIESCHANGE, col = rainbow(3))
> boxplot(Q1$FATALITIESCHANGE ~ Q1$INCREASe)
```





2) The dataset concerning hepatitis are provided in the link below
https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data

a) Import the data in R

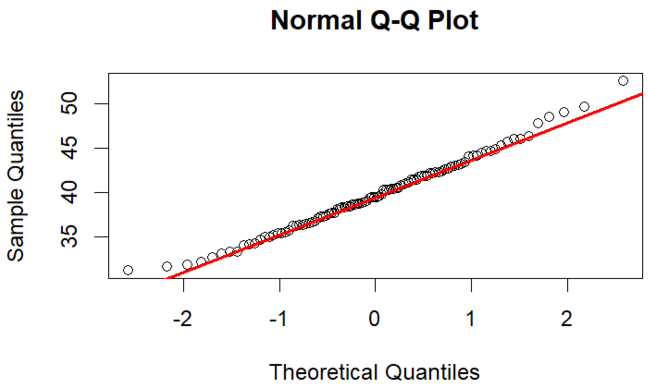
b) Missing values are marked as “? “ Replace them with NA and delete them.

c) How many observations contain missing information?

```
> url <- "https://archive.ics.uci.edu/ml/machine-learning-databases/hepatitis/hepatitis.data"
> Q2 <- read.csv(url, na.strings = "?", header = F)
> # Also, Q2[Q2 == ?] <- NA
> # Note: In read.csv, by default the first row is header
> head(Q2)
  V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20
1  2 30  2  1  2  2  2  2  1   2   2   2   2   2 1.0  85  18 4.0  NA   1
2  2 50  1  1  2  1  2  2  1   2   2   2   2   2 0.9 135  42 3.5  NA   1
3  2 78  1  2  2  1  2  2  2   2   2   2   2   2 0.7  96  32 4.0  NA   1
4  2 31  1 NA  1  2  2  2  2   2   2   2   2   2 0.7  46  52 4.0  80   1
5  2 34  1  2  2  2  2  2  2   2   2   2   2   2 1.0  NA 200 4.0  NA   1
6  2 34  1  2  2  2  2  2  2   2   2   2   2   2 0.9  95  28 4.0  75   1
> newdata <- na.omit(Q2)
> dim(newdata)
[1] 80 20
> dim(Q2)
[1] 155  20
> cat("Thus there are 75 rows with NA values")
Thus there are 75 rows with NA values
```

3) Generate 100 random numbers from a normal distribution with mean 40 and standard deviation 5. Draw the normal qq-plots .

```
> Q3 = rnorm(100, mean = 40, sd = 5)
> # q3 = rnorm(100,40,5)
> qqnorm(Q3)
> qqline(Q3, col = "red", lwd = 2)
```

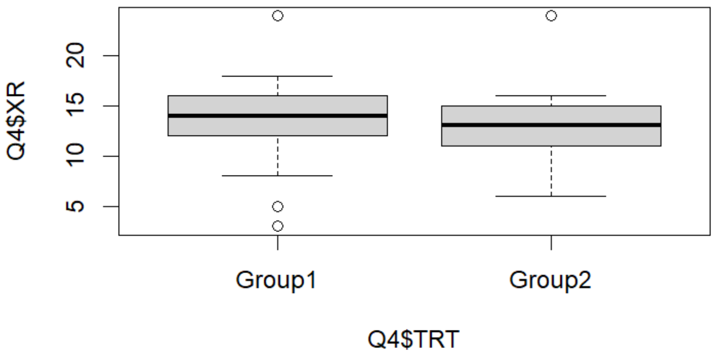


4) Data of the manuscript 'Analysis of data with censored initiating and terminating times: a missing-data approach' by Xin Tu are provided in the link below
http://lib.stat.cmu.edu/jcgs/tu

a) Import the data in R without saving in your computer and determine its dimension.

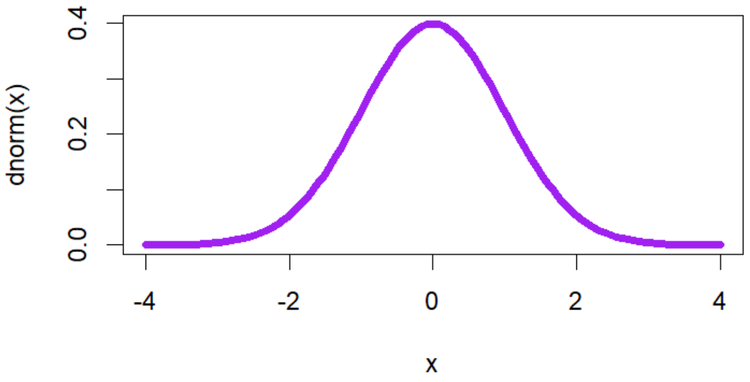
b) Display the distribution of XR values based on different treatment type (TRT).

```
> Q4 <- read.table("https://lib.stat.cmu.edu/jcgs/tu", skip = 3, header = T)
> head(Q4)
  XL XR ZL ZR AGE MULT TRT
1 15 24  1 24   1   13   1
2 15 24  1 24   2    6   1
3 16 24  1 24   1   15   1
4 16 24  1 24   2   16   1
5 17 24  1 24   1    9   1
6 17 24  1 24   2    1   1
> dim(Q4)
[1] 136   7
> boxplot(Q4$XR~Q4$TRT, names =c("Group1","Group2"))$out
 [1] 24 24 24 24 24 24 24  5  3 24 24 24 24 24
```



5) Plot the pdf of a standard normal distribution by generating data in (-4, 4).

```
> curve(dnorm(x),-4,4,col = "purple",lwd = 5)
```



6) The link below includes the crime rates for 50 states in 2005.
http://datasets.flowingdata.com/crimeRatesByState2005.tsv

a) Import the dataset in R and name it *crime*.

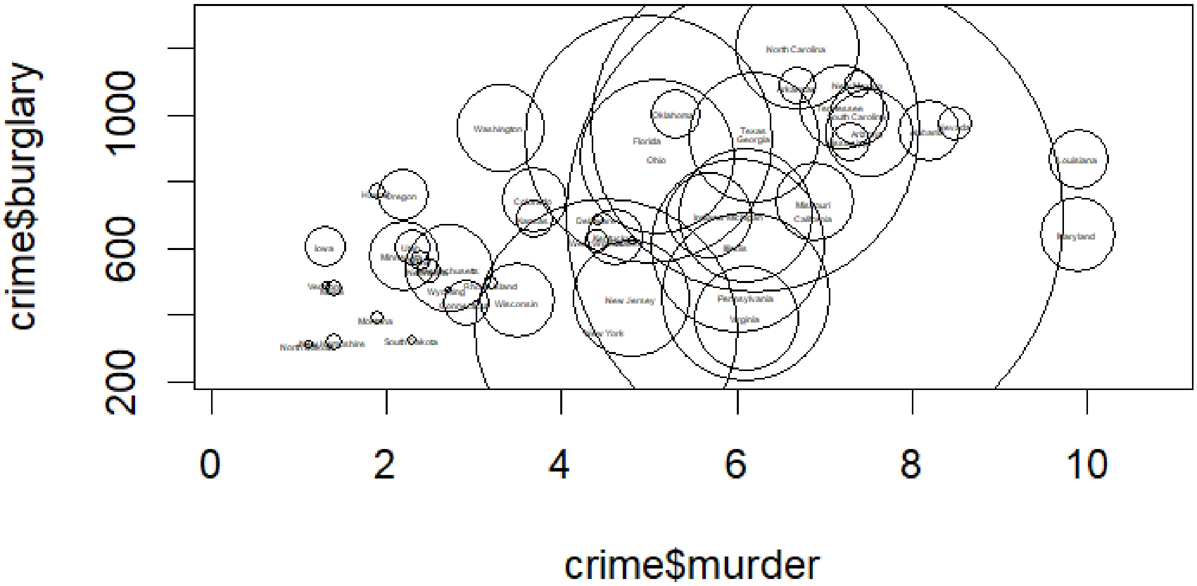
b) How many variables are included in the data

c) Use code below to draw bubble plots
symbols(crime$murder, crime$burglary, circles=crime$population)

d) Add the name of the states using code:
text(crime$murder, crime$burglary, crime$state, cex=0.5)
You may add option: bg=”red” etc.

```
        > crime <- read.table("C:\\Users\\PNW_checkout\\Downloads\\vaishak\\PNW_COURSE-WORK\\FAL
L24\\STATISTICAL COMPUTING\\Assignment\\Assignment 7\\crimeRatesByState2005.tsv", sep = "\t", hea
der = T )
> head(crime)
        state murder Forcible_rate Robbery aggravated_assult burglary larceny_theft motor_vehicle
_theft
1    Alabama     8.2          34.3   141.4             247.8    953.8        2650.0
288.3
2     Alaska     4.8          81.1    80.9             465.1    622.5        2599.1
391.0
3    Arizona     7.5          33.8   144.4             327.4    948.4        2965.2
924.4
4    Arkansas    6.7          42.9    91.1             386.8   1084.6        2711.2
262.1
5 California     6.9          26.0   176.1             317.3    693.3        1916.5
712.8
```

```
6   Colorado     3.7          43.4    84.6              264.7    744.8        2735.2
559.5
  population
1    4627851
2     686293
3    6500180
4    2855390
5   36756666
6    4861515
> variable.names(crime) # not yet counted
[1] "state"                "murder"               "Forcible_rate"        "Robbery"
[5] "aggravated_assult"    "burglary"             "larceny_theft"        "motor_vehicle_theft"
[9] "population"
> attach(crime)
> symbols(crime$murder, crime$burglary, circles = crime$population)
> text(crime$murder, crime$burglary, crime$state, cex=0.2,bg ="red")
```


crime$burglary
crime$murder
North Carolina
Arkansas
New Mexico
Tennessee
South Carolina
Arizona
Alabama
Nevada
Oklahoma
Washington
Florida
Ohio
Texas
Georgia
Louisiana
Oregon
Colorado
Delaware
Indiana
Michigan
Missouri
California
Illinois
Maryland
Iowa
Utah
Minnesota
Massachusetts
Rhode Island
Wyoming
Connecticut
Wisconsin
New Jersey
Pennsylvania
Virginia
New York
Montana
South Dakota
North Dakota
New Hampshire