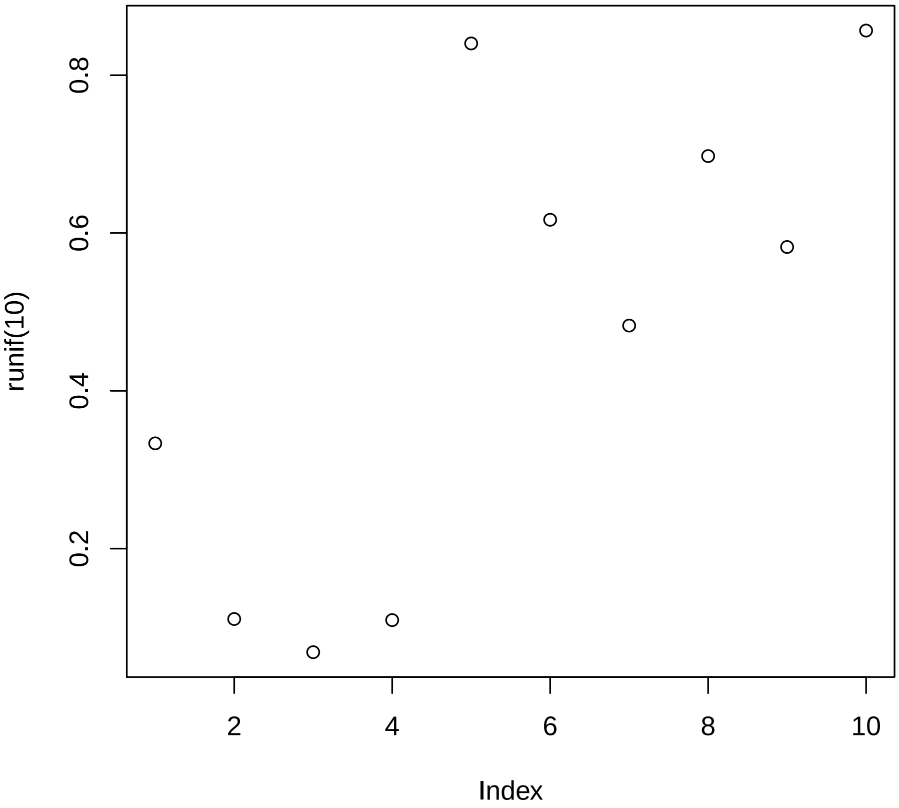
September 6th, 2019 **Solution**

Statistics I - Midterm I - Semester I 19/20

1. Consider the below outputs generated in R.

   (a) Please write the down the R command that will provide the below plot. Describe in detail what the points in the plot represent.



Solution: `runif(n, min=0, max=1)` is an imbuilt function in R that generates n random samples from the uniform distribution on the interval from 'min' to 'max'. If 'min' or 'max' are not specified they assume the default values of '0' and '1' respectively.

The above plot is executed in R by:

```
> plot(runif(10))
```

function. It is a simple scatter plot of 10 samples generated by `runif(10)`. The x-axis are the index set $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ of the samples generated and the y-axis are the corresponding sample generated.

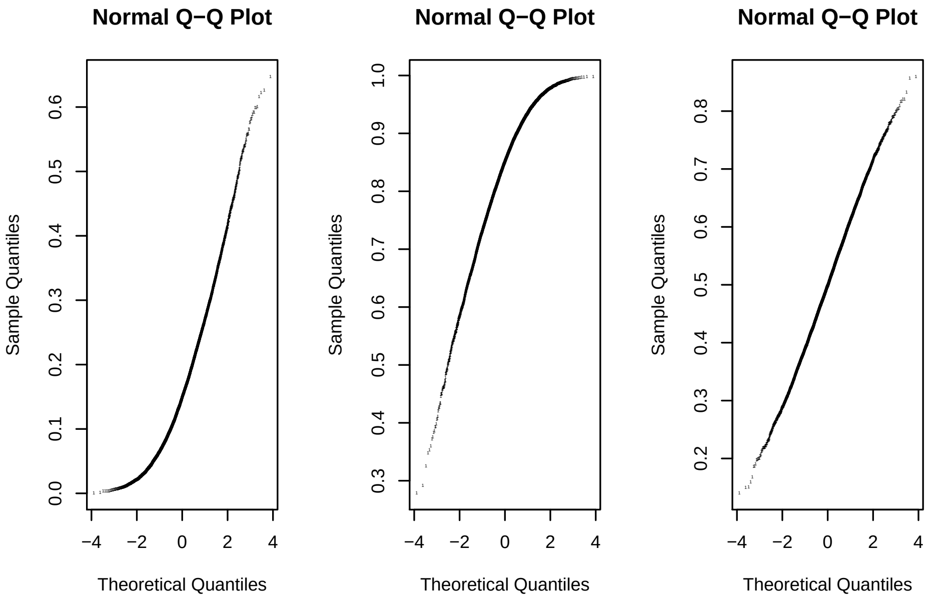
(b) Suppose the below commands are entered in R as given. Please fill in the blanks appropriately:

```
> x = c (7,-12,9,15,NA,-8,14,NA)
> x

_______________________

> which(is.na(x))

_______________________

> sum(!is.na(x))

_______________________

> x >0

_______________________
```

Solution:

```
> x = c (7,-12,9,15,NA,-8,14,NA)
> x


[1]   7 -12   9  15  NA  -8  14  NA


> which(is.na(x))


[1] 5 8

> sum(!is.na(x))


[1] 6


> x >0


[1]  TRUE FALSE  TRUE  TRUE    NA FALSE  TRUE    NA
```

(c) (5 Points) Describe the plots below. With suitable justification decide (if possible) whether the data distribution is skewed Left, Right or Symmetric.



Solution: If Z is standard Normal random variable then the α-th quantile of Z is denoted by z_α where

$$P(Z \leq z_\alpha) = \alpha, \qquad 0 < \alpha < 1.$$

Let $\{x_i : 1 \leq i \leq n\}$ be given data set then we can order them to get

$$(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$$

we view $x_{(k)}$ as the $\frac{k}{n+1}$ sample quantile. Normal Q-Q plot is a scatter plot of

$$\{(z_{\frac{k}{n+1}}, x_{(k)}) : 1 \leq k \leq n\}$$

The Normal distribution is symmetric distribution.

From the first plot we observe that the sample quantiles are less than theoretical quantiles and the slope of the curve is increasing. It indicates that the right tail is longer. Hence the distribution is likely to be Skewed Right.

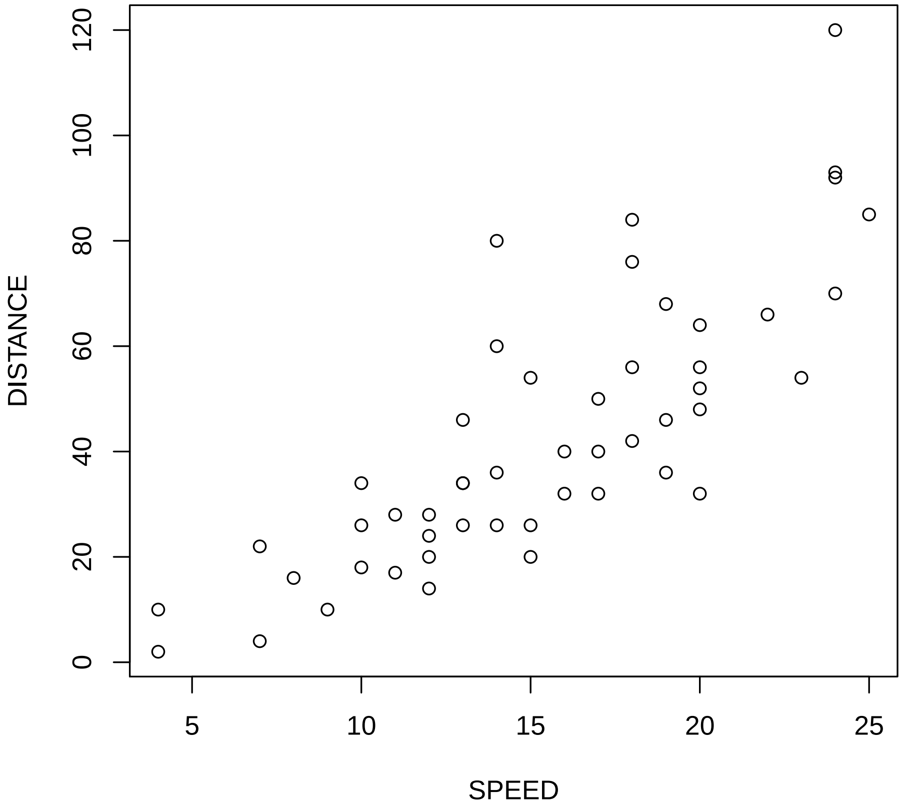
From the second plot we observe that the sample quantiles are greater than theoretical quantiles and the slope of the curve is decreasing. It indicates that the left tail is longer. Hence the distribution is Skewed left.

From the third plot we observe that the sample quantiles match the theoretical quantiles by an large. Hence the distribution is symmetric.

2. Sampagni Car company has made its latest COOL-X above road surface car. Below is a scatter plot of Distance taken to Stop versus Speed of a car taken 50 times.

(a) (5 points) Using the output of R below,

```
Call:
lm(formula = DISTANCE ~ SPEED)

Coefficients:
(Intercept)        SPEED
    -17.579        3.932
```

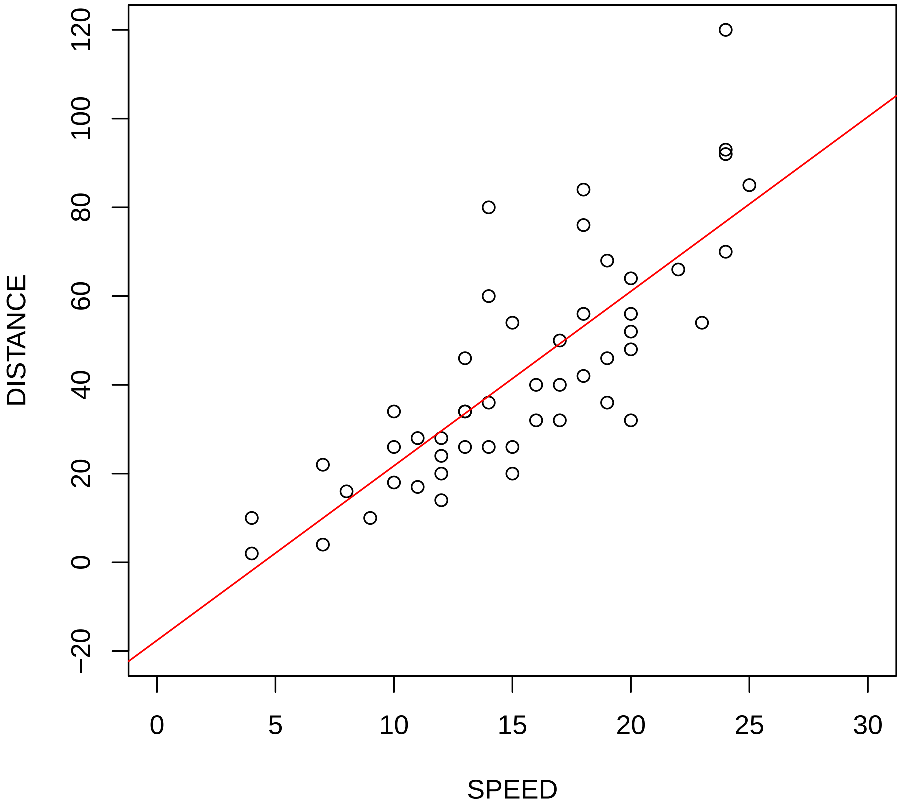


on scatter-plot draw the best possible linear relationship for the above data.

Solution:

```
> SPEED = cars$speed
> DISTANCE = cars$dist
> lm(DISTANCE~SPEED)
```

```
Call:
lm(formula = DISTANCE ~ SPEED)

Coefficients:
(Intercept)        SPEED
    -17.579        3.932

> plot(DISTANCE~SPEED, ylim=c(-20,120), xlim = c(0,30))
> abline(lm(DISTANCE~SPEED), col="red")
```



The line is $DISTANCE = 3.932(SPEED) - 17.579$.

(b) (8 points) Suppose we denote the 50 data points as

$$\{(y_i, x_i) : 1 \leq i \leq 50\}$$

Describe the output of the R command in part(a) in terms of the data points.

Solution: We wish to minimize residual sum of squares. That is find β_0, β_1 such that

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized. This can be solved using Calculus or Linear Algebra. The resulting estimates are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \bar{x},$$

where $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ The output given by R in part(a) implies that the best line fit will have slope given by : $\hat{\beta}_1 = 3.932$ and intercept given by : $\hat{\beta}_0 = -17.579$. In terms of the data:

$$3.932 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad -17.597 = \bar{y} - \left(\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \bar{x}$$

(c) (2 points) Predict the distance taken to stop when the speed is 10000

Solution: Distance take to stop = .932(10000) − 17.579 = 39302.421

3.(a) (6 points) B.Math(hons.) second year class are trying to determine the ideal amount of coffee to be sold at the canteen. They randomly sample 8 students in the canteen and ask them how much *ml* of coffee they would like to drink. The data is shown below.

$$300, 500, 300, 700, 300, 500, 300, 700.$$

Assuming normality and variancte to be 100, construct a 95% confidence interval for the true population mean of the preferred drinking amount using the above data.

Solution: For a given sample $\{x_i : 1\ leqi \leq n\}$ from a population with known variance σ^2, the 95% confidence interval is given by

$$\left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{x}, \frac{1.96\sigma}{\sqrt{n}} + \bar{x} \right)$$

Here $n = 8$, $\sigma = 10$ and

$$\bar{x} = \frac{1}{8}(300 + 500 + 300 + 700 + 300 + 500 + 300 + 700) = \frac{3600}{8} = 450.$$

Therefore the required 95% confidence interval is

$$\left(-\frac{1.96 \cdot 10}{\sqrt{8}} + 450, \frac{1.96 \cdot 10}{\sqrt{8}} + 450 \right)$$

(b) (4 points) Below is an R function intending to find the 95% confidence interval for true mean from data x with known variance 1. Please fill in the blanks.

```
cifn = function(x) {

z = qnorm( ________________________)

sdx = ________________________

c(mean(x) - z*sdx, mean(x) + z*sdx)

}
```

Solution:

```
> cifn = function(x){
+ z = qnorm(0.025, lower.tail=FALSE)
+ sdx = sqrt(1/length(x))
+ c(mean(x) - z*sdx, mean(x) + z*sdx)
+ }
```

(c) (5 points) If `cifn` is above, then please explain the intention of the below program and the output given.

```
> normaldata = replicate(100, rnorm(100,0,1),
+ simplify=FALSE)
> cidata = sapply(normaldata, cifn)
> TRUEIN = cidata[1,]*cidata[2,]<0
> table(TRUEIN)



TRUEIN
FALSE  TRUE
    3    97
```

Solution:

```
> normaldata = replicate(100, rnorm(100,0,1))
```

`rnorm(100,0,1)` generates 100 samples from Normal distribution with mean 0 and 1. `replicate(100, rnorm(100,0,1))` performs 100 repeated evaluations of `rnorm(100,0,1)`

```
> simplify=FALSE
```

ensures that result should not be simplified to a vector.

```
> cidata = sapply(normaldata, cifn)
```

returns two values giving the left end point and right end point of the 95% confidence interval for each corresponding replication of 100 samples from Normal distribution with mean 0 and variance 1.

```
> TRUEIN = cidata[1,]*cidata[2,]<0
```

returns for each of the 100 replications TRUE (if the product of the end points are negative) and FALSE otherwise.

```
> table(TRUEIN)
```

Tabulates the `TRUEIN` data.

Thus, the program provides a simulation of 100 confidence intervals derived from simulated 100 samples from Normal distribution with mean 0 and Variance 1. It also tabulates how many of them contain the true mean 0.

4. Siva's class in 1993 of 15 received the following scores out of 50:

$$21, 22, 25, 28, 29, 30, 32, 33, 33, 34, 36, 40, 41, 41, 43.$$

(a) (5 points) Find the five number summary for the data.

(b) (2 points) What is the intraquartile range of the data ?

(c) (2 points) Are there any outliers in the dataset ? Explain.

(d) (4 points) Draw a boxplot for the data.

(e) (2 points) Describe if the plot is (best option amongst) left skewed, right skewed, symmetric.

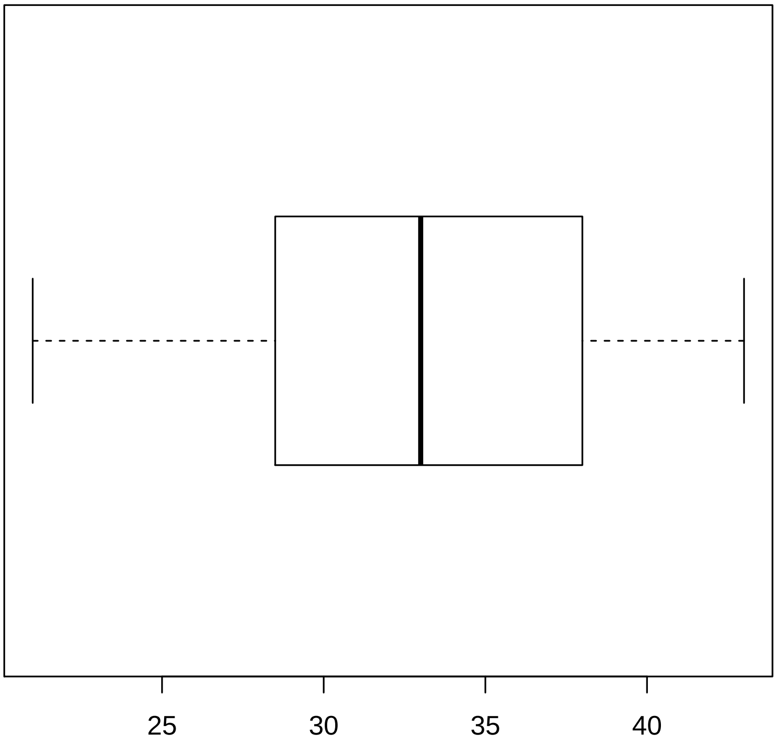
Solution: (a)

```
0%          25%          50%          75%          100%
21           28           33           40           43
```

(b) IQR = Q3-Q1 = 40-28 = 12.

(c) Lower limit for outlier $= 28 - 1.5 \times 12 = 10$. Upper limit for outlier $= 40 - 1.5 \times 12 = 58$. There are no points less than 10 and bigger than 58. Hence there are no outliers.

(d)



(e) Mode is 33. The left tail (7 points) is a bit longer than the right tail (6 points). Hence the data is left-skewed.