INTRODUCTION TO STATISTICS

PS06

Q1. In the Powerball lottery, there are 59 white balls, numbered 1 to 59. Each week, five of the white balls are drawn, without replacement. In the past, the most frequently occurring white ball has been 23.

Q1. no. of balls in the Powerball lottery = 59

given,

they are numbered from 1 to 59.

Each week, 5 of the white balls are drawn without replacement.

(a) In the next lottery, will the probability of drawing the number 23 be greater than 5/59,
less than 5/59, or equal to 5/59? Before one season, the Oakland A's were considered to be an average major league baseball team, predicted to win half (81) of their 162 games. They win the first six games of the season.

(a) Show that probability of drawing any specific number b/w 1 and 59 is 5/59.

we know

Total outcomes = 59

$P(\text{1st draw to be a specific number}) = \frac{1}{59}$

$P(\text{1st draw is not that specific number}) = \frac{1}{59}$

Favourable = $5C_1$ that is any number among 5 chosen.

$$\therefore \quad P = \frac{5C_1}{59C_1} = \frac{\frac{5!}{1!\,4!}}{\frac{59!}{1!\,58!}} = \frac{5}{59}$$

Monisha Patro

(b) True or false: After the first six games, the best prediction of the total number of games the Oakland A's win that season is 162 out of 162.

(b) given,
before one season, Oakland A's → major league champions → won 81 of 162 games.
they win first six games of the season.

Oakland A's are predicted to win half of their games : $P(\text{win}) = 0.5$

Probability that they would win 6 games in a row $= (0.5)^6$

Probability that they would win the rest 156 games $= (0.5)^{156}$

Predicted they would win 81 games $= (0.5)^{81}$
combining these probabilities
$(0.5)^6 \times (0.5)^{156} = (0.5)^{162}$
this number is very small even if their winning is slightly higher after winning the first 6 games.
Hence, FALSE

(c) True or false: After the first six games, the best prediction of the total number of games the Oakland A's win that season is still 81 out of 162.I survey a simple random sample of 1000 U.S. households and find out their income.

(c) True. Based on the probability of winning 81 of 162 game, winning the first 6 matches is the best outcome.
Hence the probability of them winning is 81/162 games still holds true.

(d) True or false: By the Central Limit Theorem, the incomes in the population will have an approximately normal distribution.

(d) True. CLT in this context means that distribution of the sample mean will be approximately normal, given the sample size is large enough i.e., 1000 in this case.

(e) True or false: By the Central Limit Theorem, the incomes in the sample will have an approximately normal distribution.

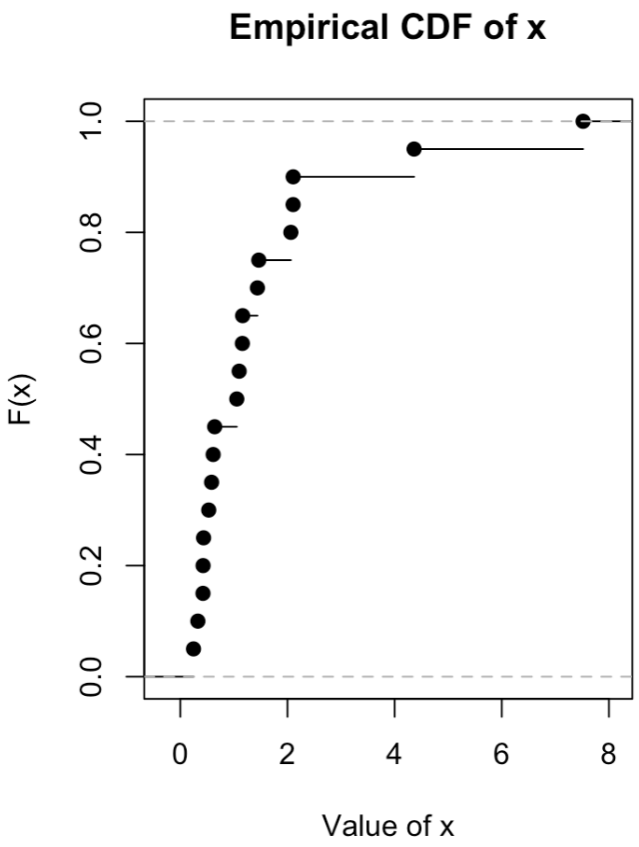
(e) False. CLT is applicable on the whole sample (the sample mean) not on individual observations.

Q2. The following sample, x , was observed and sorted:

0.246	0.327	0.423	0.425	0.434
0.530	0.583	0.613	0.641	1.054
1.098	1.158	1.163	1.439	1.464
2.063	2.105	2.106	4.363	7.517

(a) Graph the empirical cdf of x .

```
> x <- c(0.246, 0.327, 0.423, 0.425, 0.434,
+        0.530, 0.583, 0.613, 0.641, 1.054,
+        1.098, 1.158, 1.163, 1.439, 1.464,
+        2.063, 2.105, 2.106, 4.363, 7.517)
> ecdf_x <- ecdf(x)
> plot(ecdf_x, main="Empirical CDF of x", xlab="Value of x", ylab="F(x)")
> x <- c(0.246, 0.327, 0.423, 0.425, 0.434,
+        0.530, 0.583, 0.613, 0.641, 1.054,
+        1.098, 1.158, 1.163, 1.439, 1.464,
+        2.063, 2.105, 2.106, 4.363, 7.517)
> ecdf_x <- ecdf(x)
> ecdf_x
Empirical CDF
Call: ecdf(x)
 x[1:20] =  0.246,  0.327,  0.423,  ...,  4.363,  7.517
```



(b) Calculate the plug-in estimates of the mean, the variance, median, and the interquartile range.

```
> x <- c(0.246, 0.327, 0.423, 0.425, 0.434,
+        0.530, 0.583, 0.613, 0.641, 1.054,
+        1.098, 1.158, 1.163, 1.439, 1.464,
+        2.063, 2.105, 2.106, 4.363, 7.517)
> mean_x <- mean(x)
> var_x <- var(x)
> median_x <- median(x)
> iqr_x <- IQR(x)
> print(paste("Mean: ", mean_x))
[1] "Mean:  1.4876"
> print(paste("Variance: ", var_x))
[1] "Variance:  2.9342672"
> print(paste("Median: ", median_x))
[1] "Median:  1.076"
> print(paste("IQR: ", iqr_x))
[1] "IQR:  1.10775"
```

(c) Take the square root of the plug-in estimate of the variance and compare it to the plug-in estimate of the interquartile range. Do you think that x was drawn from a normal distribution? Why or why not?

The square root of the variance is the standard deviation.
Standard Deviation = √Variance
The SD provides a measure of the amount of variation or dispersion of the set of values.

IQR:
The IQR measures the statistical spread between the 25th percentile (Q1) and the 75th percentile (Q3). It's robust against outliers and gives a good sense of the spread of the central 50% of the data.

Comparison:
For a perfectly normal distribution:
IQR = 1.349 X Standard Deviation.
The 1.349 value is derived from the quantiles of the standard normal distribution.

By comparing the calculated IQR with 1.349 times the standard deviation, we can gauge how close the sample might be to a normal distribution.

```
> # Given sample
> x <- c(0.246, 0.327, 0.423, 0.425, 0.434,
+        0.530, 0.583, 0.613, 0.641, 1.054,
+        1.098, 1.158, 1.163, 1.439, 1.464,
+        2.063, 2.105, 2.106, 4.363, 7.517)
>
> # Calculations
> standard_deviation <- sqrt(var(x))
> standard_deviation
[1] 1.71297
> iqr_x <- IQR(x)
> iqr_x
[1] 1.10775
> comparison_value <- 1.349 * standard_deviation
> comparison_value
[1] 2.310797
```

Given:

Standard Deviation (SD) = 1.712

IQR = 1.107

Comparison Value = 1.349 × SD = 2.310

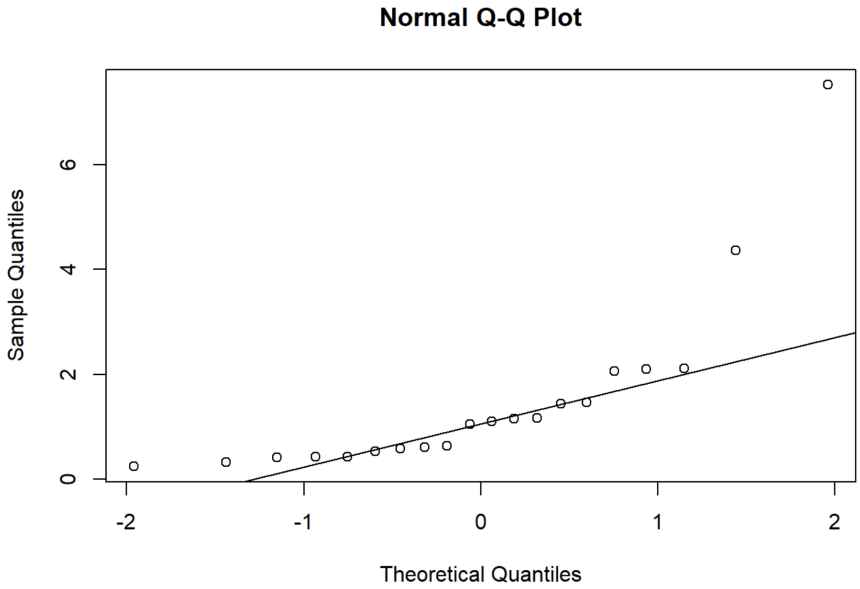
Monisha Patro

Now, let's compare the IQR with the comparison value.

Your calculated IQR is 1.107, which is significantly different from the comparison value of 2.310.

When the IQR is close to 1.349 × Standard Deviation ;1.349×Standard Deviation, it suggests that the sample might be from a normal distribution. Conversely, if they're quite different, it suggests the data may not be normally distributed.

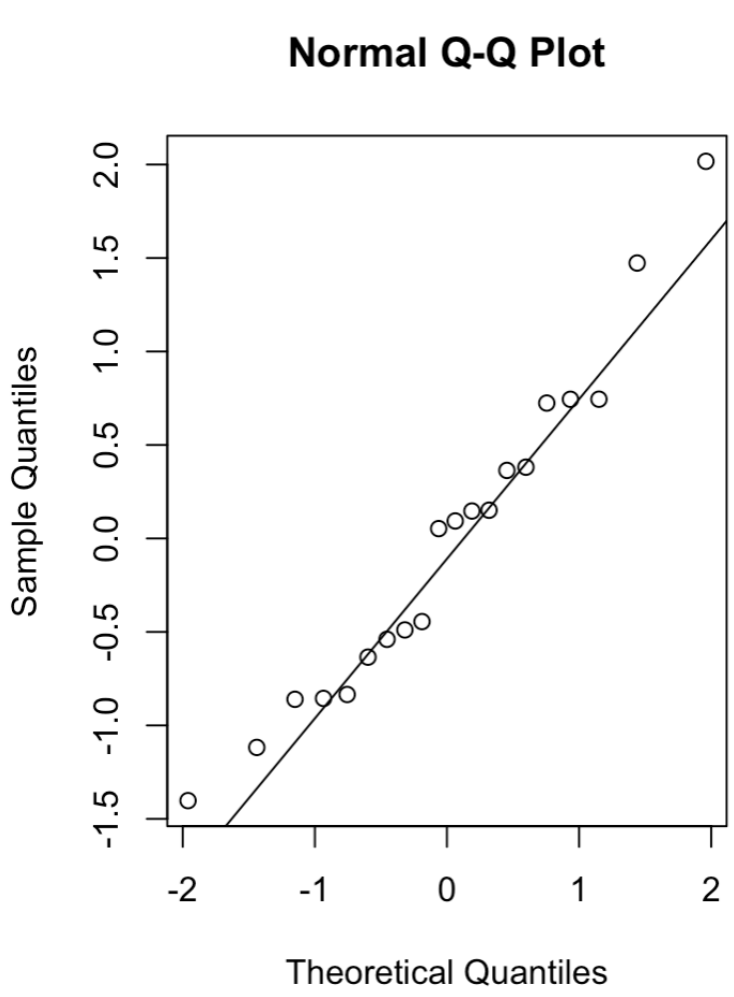
In this case, given the disparity between the IQR and the comparison value, the sample might not be from a normal distribution.



(d) Use the qqnorm function to create a normal probability plot. Do you think that x was drawn from a normal distribution? Why or why not?

```
> x <- c(0.246, 0.327, 0.423, 0.425, 0.434,
+        0.530, 0.583, 0.613, 0.641, 1.054,
+        1.098, 1.158, 1.163, 1.439, 1.464,
+        2.063, 2.105, 2.106, 4.363, 7.517)
> y <- log(x)
> y
 [1] -1.40242374 -1.11779511 -0.86038310 -0.85566611
 [5] -0.83471074 -0.63487827 -0.53956809 -0.48939034
 [9] -0.44472582  0.05259245  0.09349034  0.14669438
[13]  0.15100287  0.36394843  0.38117242  0.72416123
[17]  0.74431547  0.74479041  1.47315989  2.01716712
> qqnorm(y)
> qqline(y)
```

Monisha Patro



(e)Now consider the transformed sample y produced by replacing each x i with its natural logarithm. If x is stored in the vector x , then y can be computed by the following R command: > y <- log(x)
Do you think that y was drawn from a normal distribution? Why or why not?
As we can see, Y looks like a step function even though it has a normal distribution like trend. So we can say that Y has not been drawn from a normal distribution.

Q3. Consider an urn that contains 10 tickets, labeled { 1 , 1 , 1 , 1 , 2 , 5 , 5 , 10 , 10 , 10 } . From this urn, I propose to draw (with replacement) n = 40 tickets. I am interested in the sum, Y , of the 40 ticket values that I draw.

(a) Write an R function named urn.model that simulates this experiment, i.e evaluating urn.model is like observing a value, y , of the random variable Y .

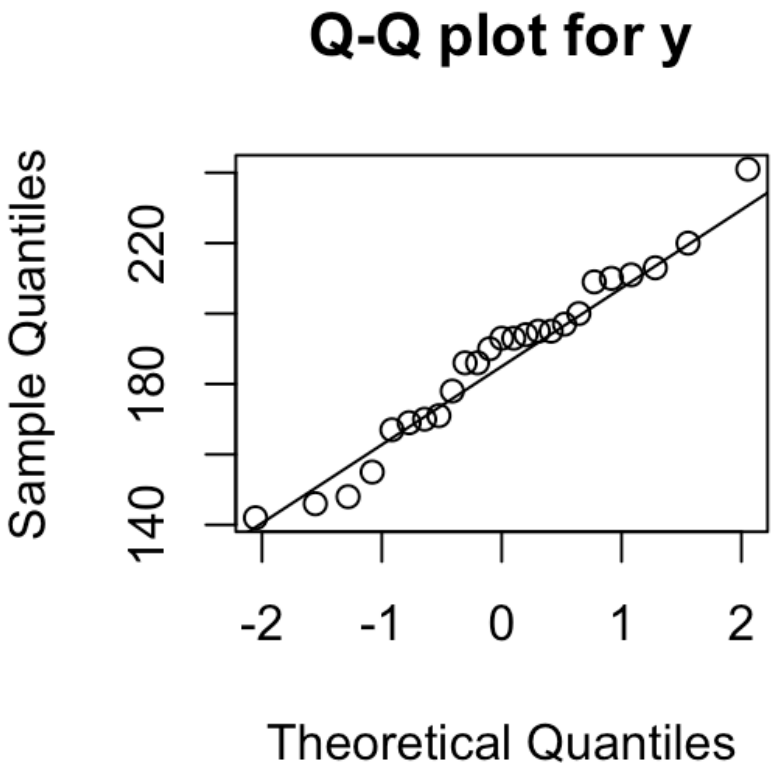
```
> #MonishaPatro
> urn.model <- function(x) {
+   # Defining the urn
+   urn <- c(1, 1, 1, 1, 2, 5, 5, 10, 10, 10)
+
+   # Draw 40 tickets with replacement
+   draws <- sample(urn, size = 40, replace = TRUE)
+   sample=sample(c(1,2,5,10),size=x)
+   x=sum(sample)
+   x
+   prob=c(4/10,1/10,2/10,3/10)
+   # Return the sum of the draws
+   return(sum(draws))
+ }
>
> # Test the function
> set.seed(124)
> y <- urn.model()
> print(y)
[1] 186
```

(b) Use urn.model to generate a sample, y = { y 1 , . . . , y 25 } , of n = 25 observed sums. The random variable Y is discrete. Does it appear that the distribution of Y can be approximated by a normal distribution? Why or why not?

```
> urn.model=function(x)
+ {
+   set.seed(124)
+   samples=sample(c(1,2,5,10),size=x,replace=TRUE, prob=c(4/10,1/10,2/10,3/10))
+   samples
+   x=sum(samples)
+   x
+ }
> urn.model(x=25)
[1] 131
>
> # 1. Histogram and Density Plot
> hist(y, prob = TRUE, main = "Histogram of y with Density", xlab = "Sum of 40 draws")
> lines(density(y), col = "blue")
>
> # 2. Q-Q Plot
> qqnorm(y, main = "Q-Q plot for y")
> qqline(y)
>
> # 3. Shapiro-Wilk Test
> test <- shapiro.test(y)
> print(test)

        Shapiro-Wilk normality test

data:  y
W = 0.97124, p-value = 0.6767
```



From the Q-Q plot, it appears that the distribution of Y closely resembles a normal distribution. It's not unusual to see slight deviations, particularly in the tails, especially when dealing with a smaller sample size. Observing that most data points are consistent with the reference line, it can be inferred that the sample's distribution might be suitably represented by a **normal distribution.**

Q4. Suppose that I toss a fair coin 100 times and observe 60 Heads . Now I decide to toss the same coin another 100 times. Does the Law of Averages imply that I should expect to observe another 40 Heads.

Monisha Patro

Q4. given
a fair coin 100 times
observe 60 heads
because there are two equally likely outcomes - heads and tails - on each given toss of a fair coin, the chance of recieving a head is 0.5.
Probability of 50 heads while tossing a fair coin 100 times is 100×0.5. this is just an average / predicted value.

There is no guarantee that every time you throw a coin, you will exactly land on exactly 50 heads. It is possible to get a large range of results around the expected number because coin tosses are inherently unpredictable.

It is not a given that you will get exactly 40 heads in the following 100 throws if you have seen 60 heads previously in first 100 tosses.

The probability of recieving a head remains 0.5 on each toss regardless of what happened on previous tosses, and each coin toss is independent of preceding tosses.
the second set of 100 throws could result in even more or fewer heads than 40.

In conclusion, there is NO 'law of averages' that predicts a certain result from a sequence of arbitrary events. Although individual results can vary greatly due to chance, anticipated values and probabilities give a general sense of what might occur over a large number of trials.

Monisha Patro

Q5. Chris owns a laser pointer that is powered by two AAAA batteries. A pair of batteries will power the pointer for an average of five hours use, with a standard deviation of 30 minutes. Chris decides to take advantage of a sale and buys 20 2-packs of AAAA batteries. What is the probability that he will get to use his laser pointer for at least 105 hours before he needs to buy more batteries?
Assume that twenty pairs of batteries is enough for the Central Limit Theorem to approximately hold.

Q5. assume X to be the random variable that represents hours taken by a pair of battery to power the pointer.

given $\mu = 5$ hours and $\sigma = 30$ minutes

converting SD to hours

$$\sigma = \frac{30}{5} = 0.5 \text{ hours}$$

Let us consider $X_1, X_2 \cdots, X_{20}$ represents 20 pairs of batteries.

consider $S = X_1 + X_2 + X_3 \cdots + X_{20}$

The mean and standard deviation of S can be computed as follows:

$$E(S) = E(X_1 + X_2 + X_3 + \cdots X_{20})$$

$$\Rightarrow E(X_1) + E(X_2) + \cdots E(X_{20})$$

$$\Rightarrow \mu + \mu + \mu \cdots + \mu = 20\mu$$

$$\Rightarrow 20(5) = 100$$

```
> pnorm(2.24,0,1)
[1] 0.9874545
```

Monisha Patro

$$\text{Var}(S) = \text{var}(X_1 + X_2 + X_3 + \cdots X_{20})$$

$$\Rightarrow \text{Var}(X_1) + \text{Var}(X_2) + \cdots \text{Var}(X_{20})$$

$$\Rightarrow \sigma^2 + \sigma^2 + \sigma^2 + \cdots \sigma^2 = 20(\sigma^2)$$

$$= 20(0.5^2) = 5$$

$$SD(S) = \sqrt{\text{Var}(S)}$$

$$= \sqrt{5} = 2.236.//$$

Now, the required probability can be calculated

$$P(S > 105) = P\left(\frac{S - \mu_S}{\sigma_S} > \frac{105 - 100}{2.236}\right)$$

$$\Rightarrow P(z > 2.24)$$

$$1 - P(z \leq 2.24)$$

$$P(z < 2.24) = 0.9875$$

$$P(S > 105) = 1 - 0.9875$$

$$= 0.0125$$

$\therefore$ Required probability $= 0.0125//.$