|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete Count |
| Results of rolling a dice | Discrete Count |
| Weight of a person | Continuous |
| Weight of Gold | Continuous |
| Distance between two places | Continuous |
| Length of a leaf | Continuous |
| Dog's weight | Continuous |
| Blue Color | Discrete Binary |
| Number of kids | Discrete Count |
| Number of tickets in Indian railways | Discrete Count |
| Number of times married | Discrete Count |
| Gender (Male or Female) | Discrete Categorical |

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Interval |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Interval |
| Years of Education | Interval |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

Solution:

Let p = Probability of getting a head in one toss = ½ = 0.5

Let q = Probability of getting a tail in one toss = 1 – p = 1 - ½ = 0.5

Thus, probability of getting two heads and one tail in three coin tosses

= p \* p \* q = 0.5 \* 0.5 \* 0.5 = 0.125

Required probability = 0.125

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1

When two dice are rolled sum can never be 1, hence probability = 0

1. Less than or equal to 4

When 2 dice are rolled total number of possible outcomes = 36

Total number of outcomes less than or equal to 4 = (1,1),(1,2),(1,3),(2,1),(2,2),(3,1) = 6

Probability of less than or equal to 4 = 6/36 = 1/6

1. Sum is divisible by 2 and 3

When 2 dice are rolled total number of possible outcomes = 36

Total number of outcomes for sum divisible by 2 and 3 = Total number of outcomes for sum divisible by 6 = (1,5),(5,1),(2,4),(4,2),(3,3) = 5

Probability of sum divisible by 2 and 3 = 5/36

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

Green balls = g = 3, Red balls = r = 2, blue balls = b = 2

With replacement : P(none of the balls is blue) = (5/7)\*(5/7) = 0.5102

Without replacement : P(none of the balls is blue) = (5/7)\*(4/6) = 0.4762

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count(X) | ProbabilityP(X) | X.P(X) |
| A | 1 | 0.015 | 0.015 |
| B | 4 | 0.20 | 0.8 |
| C | 3 | 0.65 | 1.95 |
| D | 5 | 0.005 | 0.025 |
| E | 6 | 0.01 | 0.06 |
| F | 2 | 0.120 | 0.24 |

Expected number of candies = sum(X.P(X)) = 3.09

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.



***R-code***

datacars <- read.csv('D:/SSP/Data Science ExcelR/BasicStatsLevel1 Q7.csv')

cat("\n Mean of Points = ",mean(datacars$Points), "\n")

cat("\n Mean of Score = ",mean(datacars$Score), "\n")

cat("\n Mean of Weight = ",mean(datacars$Weigh), "\n")

cat("\n Median of Points = ",median(datacars$Points), "\n")

cat("\n Median of Score = ",median(datacars$Score), "\n")

cat("\n Median of Weight = ",median(datacars$Weigh), "\n")

cat("\n Mode of Points = ",mode(datacars$Points), "\n")

cat("\n Mode of Score = ",mode(datacars$Score), "\n")

cat("\n Mode of Weight = ",mode(datacars$Weigh), "\n")

cat("\n Variance of Points = ",var(datacars$Points), "\n")

cat("\n Variance of Score = ",var(datacars$Score), "\n")

cat("\n Variance of Weight = ",var(datacars$Weigh), "\n")

cat("\n Standard Deviation of Points = ",sd(datacars$Points), "\n")

cat("\n Standard Deviation of Score = ",sd(datacars$Score), "\n")

cat("\n Standard Deviation of Weight = ",sd(datacars$Weigh), "\n")

cat("\n Range of Points = ",max(datacars$Points) - min(datacars$Points), "\n")

cat("\n Range of Score = ",max(datacars$Score) - min(datacars$Score), "\n")

cat("\n Range of Weight = ",max(datacars$Weigh) - min(datacars$Weigh), "\n")

***R-output***

Mean of Points = 3.596563

Mean of Score = 3.21725

Mean of Weight = 17.84875

Median of Points = 3.695

Median of Score = 3.325

Median of Weight = 17.71

Mode of Points = 3.92

Mode of Score = 3.44

Mode of Weight = 17.02

Variance of Points = 0.2858814

Variance of Score = 0.957379

Variance of Weight = 3.193166

Standard Deviation of Points = 0.5346787

Standard Deviation of Score = 0.9784574

Standard Deviation of Weight = 1.786943

Range of Points = 2.17

Range of Score = 3.911

Range of Weight = 8.4

***Observation :***

The variance of weights is much more as compared to Points and Score.

The data of weights is therefore has a wider spread.

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Expected Value = Mean of all weights

= 145.3333 pounds

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**



***R-code***

*datacars <- read.csv('D:/SSP/Data Science ExcelR/Basic Statistics/Assignments/Q9\_a.csv')*

*View(datacars)*

*install.packages('moments')*

*library(moments)*

*skewness(datacars)*

*kurtosis(datacars)*

***R-output***

Index speed dist

0.0000000 -0.1139548 0.7824835

*Thus,*

*Skewness for speed = -0.1139548*

*Skewness for dist = 0.7824835*

> kurtosis(datacars)

Index speed dist

1.799040 2.422853 3.248019

*Kurtosis for speed =* 2.422853

*Kurtosis for dist =* 3.248019

***Observation :***

Speed :

Skewness is -0.1139548, that is negative and not zero, hence the speed data is negatively/left skewed, that is mass of the data is to the right

Kurtosis is 2.422853 which positive hence the speed data is high peaked

Dist :

Skewness is *0.7824835*, that is positive and not zero, hence the distance data is positively/right skewed, that is mass of the data is to the left

Kurtosis is 3.248019 which positive hence the distance data is high peaked

**SP and Weight(WT)**



***R-code***

spwt <- read.csv('D:/SSP/Data Science ExcelR/Basic Statistics/Assignments/Q9\_b.csv')

View(spwt)

library(moments)

skewness(spwt)

kurtosis(spwt)

***R-output***

> skewness(spwt)

X SP WT

0.0000000 1.5814537 -0.6033099

> kurtosis(spwt)

X SP WT

1.799634 5.723521 3.819466

***Observation :***

SP :

Skewness is 1.5814537, that is positive and not zero, hence the SP data is positively/right skewed, that is mass of the data is to the left.

Kurtosis is 5.723521 which positive hence the SP data is high peaked.

WT :

Skewness is -0.6033099, that is positive and not zero, hence the WT data is negatively/left skewed, that is mass of the WT data is to the right.

Kurtosis is 3.819466 which positive hence the WT data is high peaked.

**Q10) Draw inferences about the following boxplot & histogram**



The boxplot and the histogram are positively or right skewed. That is major mass of data is to the left and the tail is longer at the right end.

This means majority of the chicks have weights on the lower side that is between 0 to 200

From the box-plot it can be seen that there are outliers on the positive side.

Also from the boxplot it can be seen that the median is on the lower side of the data value.

From the histogram we can see that Modal class is between 50 to 100 weight

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval ?

***R-code***

# Confidence interval

sampmean <- 200

sampsd <- 30

# sample size is nn

nn = 2000

# For 99% confidence interval alpha = 1 - 0.99 = 0.01

alpha = 0.01

# Calculate the critical z-score

z = qnorm(1-alpha/2)

err <- z\*sampsd/sqrt(nn)

# Compute the CI

ci <- sampmean + c(-1,1)\*err

cat("\n Sample Mean = ", sampmean,"\n")

cat("\n Sample Standard Deviation = ",sampsd,"\n")

cat("\n Lower Limit of Confidence interval ", ci[1],"\n")

cat("\n Upper Limit of Confidence interval ", ci[2],"\n")

***R-output***

Sample Mean = 200

Sample Standard Deviation = 30

Lower Limit of Confidence interval 198.2721

Upper Limit of Confidence interval 201.7279

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

***1)***

***R-code***

scores <- c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)

cat("\n Mean of Scores = ",mean(scores), "\n")

cat("\n Median of Scores = ",median(scores), "\n")

cat("\n Variance of Scores = ",var(scores), "\n")

cat("\n Standard Deviation of Scores = ",sd(scores), "\n")

***R-output***

Mean of Scores = 41

Median of Scores = 40.5

Variance of Scores = 25.52941

Standard Deviation of Scores = 5.052664

2) Since the mean is more than the median, the data appears to be right skewed implying that majority of the data is to the lower side or left side and tail is long on the right side

The data varies on an average 5.053 (standard deviation) points up or below the mean.

Q13) What is the nature of skewness when mean, median of data are equal?

When mean, median of data are equal the skewness is 0

Q14) What is the nature of skewness when mean > median ?

When mean > median the skewness is positive or right.

Q15) What is the nature of skewness when median > mean?

When median > mean the skewness is negative or left.

Q16) What does positive kurtosis value indicates for a data ?

Positive kurtosis indicates that the data is high peaked and heavy tailed.,

Q17) What does negative kurtosis value indicates for a data?

Negative kurtosis indicates a flatter curve with multiple peaks.

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

The data distribution is not symmetric or normal

What is nature of skewness of the data?

The data distribution is skewed to the left or is negatively skewed as the left tail is long.

What will be the IQR of the data (approximately)?   
  
IQR = Q3 – Q1

Q3 = 18, Q1 = 10

IQR = 18 – 10

IQR = 8  
  
Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

The median of the box plots 1 and 2 both are same at about 262

But boxplot 2 appears to be more symmetric and normal than boxplot 1

Boxplot 1 is slightly positively skewed

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)
  2. P(MPG<40)

c. P (20<MPG<50)

***R-code***

carsdata<-read.csv(file.choose())

View(MPG)

MPG <- carsdata$MPG

#P(MPG > 38)

pp <- (length(which(MPG > 38)))/length(MPG)

cat("\n P(MPG > 38) = ",pp, "\n")

#P(MPG < 40)

pp <- (length(which(MPG < 40)))/length(MPG)

cat("\n P(MPG < 40) = ",pp, "\n")

#P(20 < MPG < 50)

pp <- (length(which(MPG < 50)) - length(which(MPG < 20)))/length(MPG)

cat("\n P(20 < MPG < 50) = ",pp, "\n")

***R-output***

P(MPG > 38) = 0.4074074

P(MPG < 40) = 0.7530864

P(20 < MPG < 50) = 0.8518519

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

***R-code***

carsdata<-read.csv(file.choose())

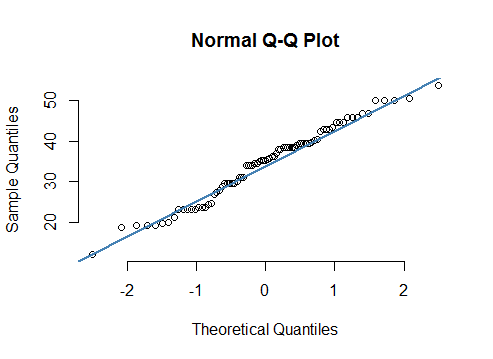
MPG <- carsdata$MPG

library(ggplot2)

qqnorm(MPG, pch = 1, frame=FALSE)

qqline(MPG, col = "steelblue", lwd = 2)

***R-output***

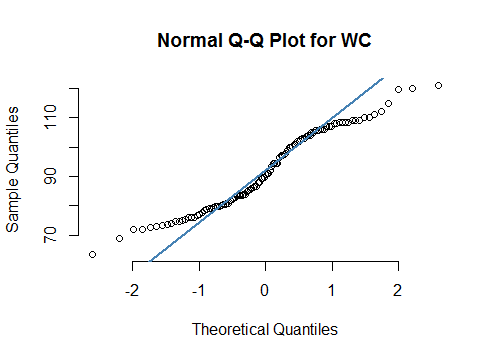


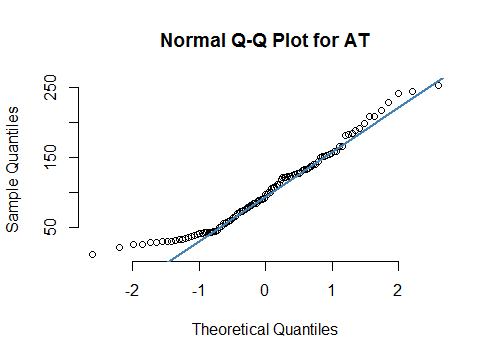
Since most of the points lie along the straight line, MPG of cars follows a Normal Distribution.

1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

***R-output***





From the QQ plots it can be seen that the waist circumference data points do not fall on a line, hence the distribution of waist circumference is not normal.

About 80% of the Adipose tissue data points fall on a line hence the distribution of the AT points is close to normal although not completely normally distributed.

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

***R-code***

z90 = qnorm(0.95)

z90

z94 = qnorm(0.93)

z94

z60 = qnorm(0.80)

z60

Z score for 90% = 1.644854

Z score for 94% = 1.475791

Z score for 60% = 0.8416212

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

***R-code***

df = 25 - 1

# t score for 95% CI

t95 = qt(0.975, df)

t95

# t score for 96% CI

t96 = qt(0.98, df)

t96

# t score for 99% CI

t99 = qt(0.995, df)

t99

t score for 95% = 2.063899

t score for 96% = 2.171545

t score for 99% = 2.79694

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode 🡪 pt(tscore,df)

df 🡪 degrees of freedom

***R-code***

df = 18 - 1

mu = 260

stdev = 90

nn = 18

tscore = (260-260)/90

tscore

# To find pp = P(x < 260)

pp = pt(0, df)

pp

***R-output***

0.5

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days = **0.5**