Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| 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 |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

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

Nominal, Ordinal, Interval, Ratio.

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

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

Ans:

Total Number of Possible Events = HHH, HHT, HTH, THH, HTT, THT, TTH, TTT

Total Number of Interested Events = HHT, HTH, THH

Probability= Total Number of Interested Events/ Total Number of Possible Events

=3/8

=0.375

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

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

Ans:

Total Number of Possible Events= 36 [(1,2),…..,(6,6)]

a) Probability= 0/ 36 = 0

b) Probability= 6/36=0.16667 [(1,1)(1,2)(1,3)(2,1)(2,2)(3,1)]

c) Probability=6/36=0.1111 [(5,1) (4,2) (3,3), (2,4), (1,5), (6,6)]

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?

Ans:

Total Number of Possible Events = 6+5+4+3+2+1 = 21

Total Number of Interested Events = 4+3+2+1 = 10

Probability= Total Number of Interested Events/ Total Number of Possible Events

=10/21

= 0.476

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

Ans:

Expected number of candies for a randomly selected child

= 1 x 0.015 + 4 x 0.20 + 3 x 0.65 + 5 x 0.005 + 6 x 0.01 + 2 x 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 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.



> Q7 <- read.csv(choose.files())

> str(Q7)

'data.frame': 32 obs. of 4 variables:

$ X : Factor w/ 32 levels "AMC Javelin",..: 18 19 5 13 14 31 7 21 20 22 ...

$ Points: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...

$ Score : num 2.62 2.88 2.32 3.21 3.44 ...

$ Weigh : num 16.5 17 18.6 19.4 17 ...

> mean(Q7$Points)

[1] 3.596563

> mean(Q7$Score)

[1] 3.21725

> mean(Q7$Weigh)

[1] 17.84875

> median(Q7$Points)

[1] 3.695

> median(Q7$Score)

[1] 3.325

> median(Q7$Weigh)

[1] 17.71

> mode(Q7$Points)

[1] "numeric"

> mode(Q7$Score)

[1] "numeric"

> mode(Q7$Weigh)

[1] "numeric"

> getmode<-function(x){

+ uniquv<-unique(x)

+ uniquv[which.max(tabulate(match(x,uniquv)))]

+ }

> getmode(Q7$Points)

[1] 3.92

> getmode(Q7$Score)

[1] 3.44

> getmode(Q7$Weigh)

[1] 17.02

> var(Q7$Points)

[1] 0.2858814

> var(Q7$Score)

[1] 0.957379

> var(Q7$Weigh)

[1] 3.193166

> sd(Q7$Points)

[1] 0.5346787

> sd(Q7$Score)

[1] 0.9784574

> sd(Q7$Weigh)

[1] 1.786943

> range(Q7$Points)

[1] 2.76 4.93

> range(Q7$Score)

[1] 1.513 5.424

> range(Q7$Weigh)

[1] 14.5 22.9

> rangevalue <- function(x){max(x)-min(x)}

> rangevalue(Q7$Points)

[1] 2.17

> rangevalue(Q7$Score)

[1] 3.911

> rangevalue(Q7$Weigh)

[1] 8.4

Inference Drawn:

* The mean is useful for spotting trends in the data because we can compare means over a time period to spot trends. The mean is the most common measure of central tendency.
* The **median** divides a sample of data in half; it is the middle score. The median is a useful statistic if we think our data have some extreme cases. The median is not impacted by extreme cases, but the mean is.

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?

Ans:

> weights<- c(108, 110, 123, 134, 135, 145, 167, 187, 199)

> mean(weights)

[1] 145.3333

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

**Cars speed and distance**



> Q9a<-read.csv(choose.files())

> str(Q9a)

'data.frame': 50 obs. of 3 variables:

$ Index: int 1 2 3 4 5 6 7 8 9 10 ...

$ speed: int 4 4 7 7 8 9 10 10 10 11 ...

$ dist : int 2 10 4 22 16 10 18 26 34 17 ...

> library(moments)

> skewness(Q9a$speed)

[1] -0.1139548

> kurtosis(Q9a$speed)

[1] 2.422853

> skewness(Q9a$dist)

[1] 0.7824835

> kurtosis(Q9a$dist)

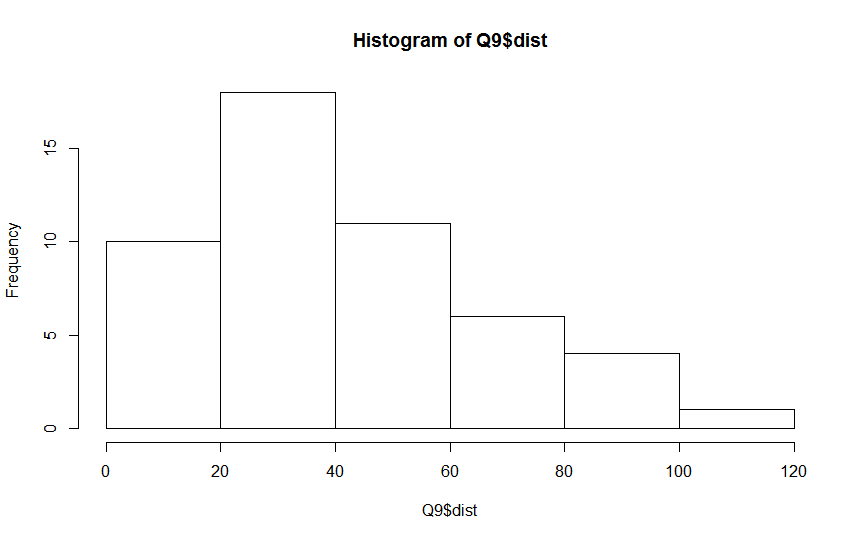
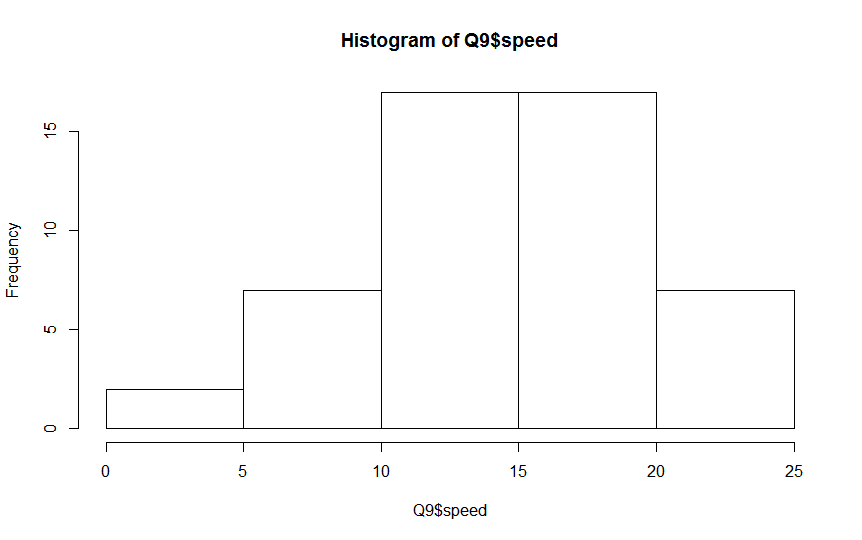
[1] 3.248019

> hist(Q9a$speed)

> hist(Q9a$dist)

**Speed: Data has negative shewness (tail is on the left side) and positive kurtosis (heavy-tailed), Mean of distribution is less than the Median.**

**Dist: Data has positive shewness (tail is on the right side) and positive kurtosis (heavy-tailed).** **Mean of distribution is more than the Median.**

****

**SP and Weight(WT)**



> Q9b<-read.csv(choose.files())

> str(Q9b)

'data.frame': 81 obs. of 3 variables:

$ X : int 1 2 3 4 5 6 7 8 9 10 ...

$ SP: num 104 105 105 113 104 ...

$ WT: num 28.8 30.5 30.2 30.6 29.9 ...

> skewness(Q9b$SP)

[1] 1.581454

> kurtosis(Q9b$SP)

[1] 5.723521

> skewness(Q9b$WT)

[1] -0.6033099

> kurtosis(Q9b$WT)

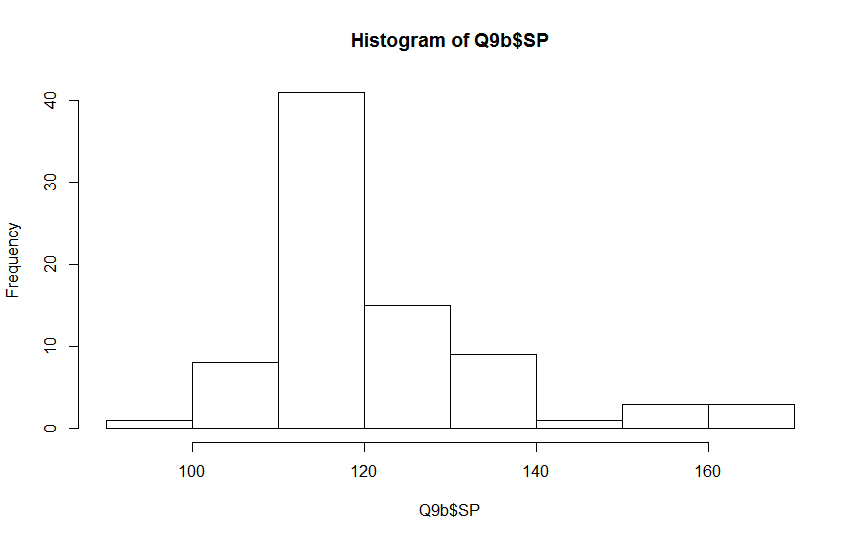
[1] 3.819466

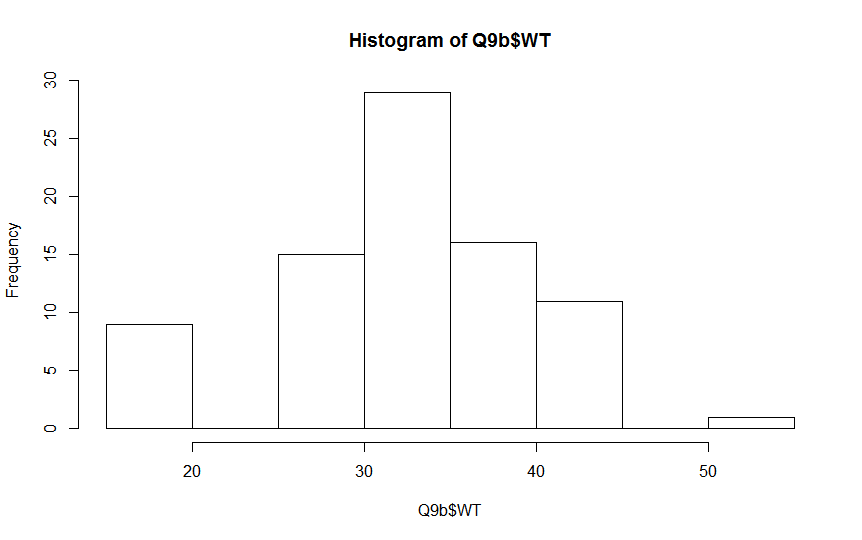
> hist(Q9b$SP)

> hist(Q9b$WT)

**SP: Data has positive shewness (long tail is on the right side) and positive kurtosis (heavy-tailed). Mean of distribution is more than the Median.**

**Dist: Data has negative shewness (long tail is on the left side) and positive kurtosis (heavy-tailed). Mean of distribution is less than the Median.**

****

****

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



**Ans:**

**Inferences: 1) Average Chick Weight – 100-150**

**2) Range of Chick Weight – 0 to 400**

**3) Data has positive skewness (long tail is towards right side)**

**4) Data has positive kurtosis (Thinner peak)**

**5) Mean > Median**

**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?

Ans:

Population (N) = 3,000,000

Sample Size (n) = 2000

Xbar = 200

Sample SD (S) = 30

Degree of Freedom (df) = 2000-1= 1999

1. 94%

> N<-3000000

> n<-2000

> xbar<-200

> s<-30

> df<-1999

> Upperlimt<-xbar+qt(0.97,1999)\*s/sqrt(n)

> print(Upperlimt)

[1] 201.2624

> Lowerlimt<-xbar-qt(0.97,1999)\*s/sqrt(n)

> print(Lowerlimt)

[1] 198.7376

1. 98%

> N<-3000000

> n<-2000

> xbar<-200

> s<-30

> df<-1999

> Upperlimt<-xbar+qt(0.99,1999)\*s/sqrt(n)

> print(Upperlimt)

[1] 201.5618

> Lowerlimt<-xbar-qt(0.99,1999)\*s/sqrt(n)

> print(Lowerlimt)

[1] 198.4382

1. 96%

> N<-3000000

> n<-2000

> xbar<-200

> s<-30

> df<-1999

> Upperlimt<-xbar+qt(0.98,1999)\*s/sqrt(n)

> print(Upperlimt)

[1] 201.3786

> Lowerlimt<-xbar-qt(0.98,1999)\*s/sqrt(n)

> print(Lowerlimt)

[1] 198.6214

**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.

Ans:

> score = c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)

> mean(score)

[1] 41

> median(score)

[1] 40.5

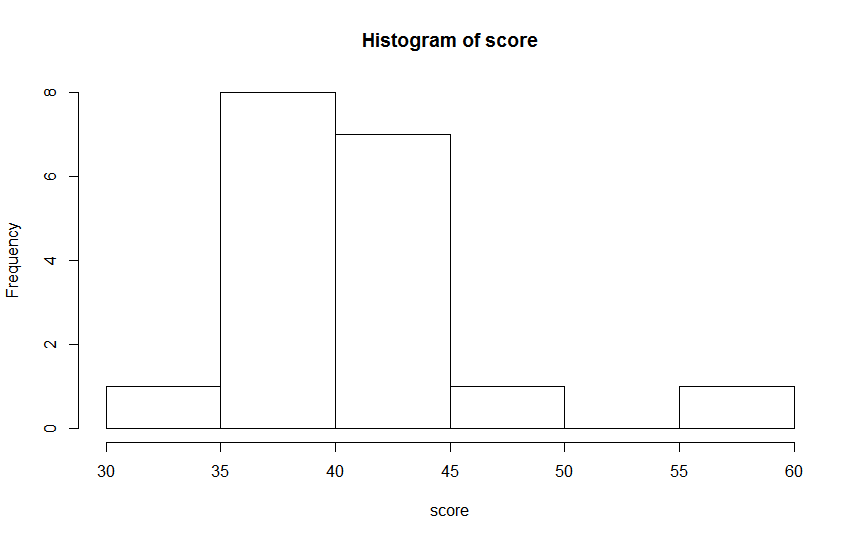
> var(score)

[1] 25.52941

> sd(score)

[1] 5.052664

1. What can we say about the student marks?



**Ans:**

**Inferences: 1) Average Student Marks – 40-45**

**2) Range of Chick Weight – 30 to 60**

**3) Data has positive skewness (long tail is towards right side)**

**4) Data has positive kurtosis (Thinner peak)**

**5) Mean > Median. No outliers are present**

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

Ans: Normal Distribution with zero skewness

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

Ans: The distribution is positively skewed

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

Ans: The distribution is negatively skewed

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

Ans: The distribution has heavier tails and a sharper peak than the normal distribution

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

Ans: The distribution has lighter tails and a flatter peak than the normal distribution

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



What can we say about the distribution of the data?

Ans: Not a Normal Distribution

What is nature of skewness of the data?

Ans: The distribution is negatively skewed

What will be the IQR of the data (approximately)?   
Ans: (10,18)=8

Q19) Comment on the below Boxplot visualizations?



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

**Ans:**

**Inferences: 1) Average of both the boxplots are same**

**2) Boxplot one has lower range as compared to boxplot 2**

**3) Both the boxplots has 0 skewness**

**4) Both the boxplots has 0 kurtosis**

**5)IQR of boxplot 1 is lower than boxplot 2**

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)

> cars<-read.csv(choose.files())

> mean<-mean(cars$MPG)

> print(mean)

[1] 34.42208

> sd<-sd(cars$MPG)

> print(sd)

[1] 9.131445

> 1-pnorm(38,mean,sd)

[1] 0.3475939

* 1. P(MPG<40)

> cars<-read.csv(choose.files())

> mean<-mean(cars$MPG)

> print(mean)

[1] 34.42208

> sd<-sd(cars$MPG)

> print(sd)

[1] 9.131445

> pnorm(40,mean,sd)

[1] 0.7293499

* 1. P (20<MPG<50)

> cars<-read.csv(choose.files())

> mean<-mean(cars$MPG)

> print(mean)

[1] 34.42208

> sd<-sd(cars$MPG)

> print(sd)

[1] 9.131445

> pnorm(50,mean,sd)-pnorm(20,mean,sd)

[1] 0.8988689

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

> cars<-read.csv(choose.files())

> shapiro.test(cars$MPG)

Shapiro-Wilk normality test

data: cars$MPG

W = 0.97797, p-value = 0.1764

Data is normally distributed as P value is high

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

Dataset: wc-at.csv

> wcat<-read.csv(choose.files())

> shapiro.test(wcat$AT)

Shapiro-Wilk normality test

data: wcat$AT

W = 0.95234, p-value = 0.000654

> shapiro.test(wcat$Waist)

Shapiro-Wilk normality test

data: wcat$Waist

W = 0.95586, p-value = 0.00117

Data is not normally distributed as P value is low

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

Ans:

> qnorm(0.95)

[1] 1.644854

> qnorm(0.97)

[1] 1.880794

> qnorm(0.8)

[1] 0.8416212

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

Ans:

> qt(0.975,24)

[1] 2.063899

> qt(0.98,24)

[1] 2.171545

> qt(0.995,24)

[1] 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

Ans:

Mu – 270

n- 18

Xbar – 260

s- 90

> mu=270

> n=18

> xbar=260

> s=90

> tscore=(xbar-mu)/(s/sqrt(n))

> pt(tscore,n-1)

[1] 0.3216725

32.16% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 260 days.