**Assignment 1 – Basic Statistics Level 1**

Q1) Identify the Data type for the Following:

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

Q2) Identify the Data types, which were among the following Nominal, Ordinal, Interval, Ratio.

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

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

**Solution:--**

When three coins are tossed the total number of possible combinations are 23 = 8.

These combinations are -HHH, HHT, HTH, THH, TTH, THT, HTT, TTT.

The number of combinations which have two heads and one tail are:

HHT, HTH, TTH which makes them 3 in number.

Therefore the Probability of getting two heads and one tails in the toss of three coins simultaneously is defined as:

P(Two heads and One Tail)

= Number of desired outcomes

= 3

= 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
4. Solution:
5. With two dice, there are ( 6 ) \* ( 6 ) = ( 36 ) possible combinations of numbers.
6. The minimum sum possible for the two dice thrown is (1, 1) = a sum of (2 )
7. The maximum sum possible for the two dice thrown is (6, 6) = a sum of (12).
8. **a)Equal to 1**
9. The minimum possible sum is (1, 1) = ( 2 ).
10. Therefore P( 1 ) = ( 0 )/( 36 ) = 0
11. **b) Less than or equal to 4**
12. The possible pairs of numbers with a sum less than or equal to 4 are (1,1), (1,2), (2,1), (2,2), (1,3), and (3,1).
13. There are 6 such pairs, so the probability of getting a sum less than or equal to 4 is = 6/36
14. = 1/6.
15. **c)Sum is divisible by 2 and 3**
16. The sum of the two numbers is divisible by 2 if and only if both numbers are even or both numbers are odd.
17. The sum of the two numbers is divisible by 3 if and only if the sum of the digits is divisible by 3.
18. Therefore, the pairs of numbers with a sum divisible by 2 and 3 are (2,4), (4,2), and (6,6), which have a sum divisible by 3, and (1,5), (3,3), and (5,1), which do not have a sum divisible by 3.
19. So there are 6 possible pairs of numbers with a sum divisible by 2 and 3, out of 36 possible pairs, giving a probability of 6/36 = 1/6.

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?

**Solution:**

Total number of balls  
= (2 + 3 + 2)  
= 7  
Let S be the sample space  
Then, n(S) = Number of ways of drawing 2 balls out of 7

n(S)=7C2

n(S)=(7\*6)/(2\*1)

n(S)=21

Let E = Event of 2 balls, none of which is blue

∴ n(E) = Number of ways of drawing 2 balls out of (2 + 3) balls

n(E)=5C2

n(E)=(5\*4)/(2\*1)

n(E)=10

P(E)=n(E)/n(S)

=10/21

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

Sol: Solution:

Expected number of candies for a randomly selected child  = 3.09

Expected number of candies for a randomly selected child

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

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

=       3.090

=  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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mazda RX4 | 3.9 | 2.62 | 16.46 |
| Mazda RX4 Wag | 3.9 | 2.875 | 17.02 |
| Datsun 710 | 3.85 | 2.32 | 18.61 |
| Hornet 4 Drive | 3.08 | 3.215 | 19.44 |
| Hornet Sportabout | 3.15 | 3.44 | 17.02 |
| Valiant | 2.76 | 3.46 | 20.22 |
| Duster 360 | 3.21 | 3.57 | 15.84 |
| Merc 240D | 3.69 | 3.19 | 20 |
| Merc 230 | 3.92 | 3.15 | 22.9 |
| Merc 280 | 3.92 | 3.44 | 18.3 |
| Merc 280C | 3.92 | 3.44 | 18.9 |
| Merc 450SE | 3.07 | 4.07 | 17.4 |
| Merc 450SL | 3.07 | 3.73 | 17.6 |
| Merc 450SLC | 3.07 | 3.78 | 18 |
| Cadillac Fleetwood | 2.93 | 5.25 | 17.98 |
| Lincoln Continental | 3 | 5.424 | 17.82 |
| Chrysler Imperial | 3.23 | 5.345 | 17.42 |
| Fiat 128 | 4.08 | 2.2 | 19.47 |
| Honda Civic | 4.93 | 1.615 | 18.52 |
| Toyota Corolla | 4.22 | 1.835 | 19.9 |
| Toyota Corona | 3.7 | 2.465 | 20.01 |
| Dodge Challenger | 2.76 | 3.52 | 16.87 |
| AMC Javelin | 3.15 | 3.435 | 17.3 |
| Camaro Z28 | 3.73 | 3.84 | 15.41 |
| Pontiac Firebird | 3.08 | 3.845 | 17.05 |
| Fiat X1-9 | 4.08 | 1.935 | 18.9 |
| Porsche 914-2 | 4.43 | 2.14 | 16.7 |
| Lotus Europa | 3.77 | 1.513 | 16.9 |
| Ford Pantera L | 4.22 | 3.17 | 14.5 |
| Ferrari Dino | 3.62 | 2.77 | 15.5 |
| Maserati Bora | 3.54 | 3.57 | 14.6 |
| Volvo 142E | 4.11 | 2.78 | 18.6 |

Sol: **Points**: Mean =3.596563, Median= 3.695, Mode= “numeric”, Variance= 0.2858814, Standard deviation= 0.5346787.

**Score:** Mean= 3.21725, Median= 3.325, Mode= “numeric”, Variance= 0.957379, Standard deviation= 0.9784574 Note: Mean value are closer for both ‘Point’ and ‘Score’.

**Weight:** Mean= 17.84875, Median= 17.71, Mode= “numeric”, Variance= 3.193166, Standard deviation= 1.786943

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

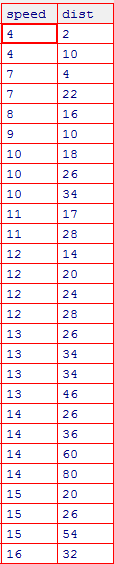
Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Sol: Mean= 1308

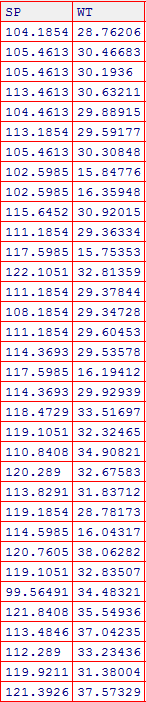
` Mue= 163.5

Expected value= ∑[x.p(x)]

# Q9) Calculate Skewness, Kurtosis & draw inferences on the following data Cars speed and distance



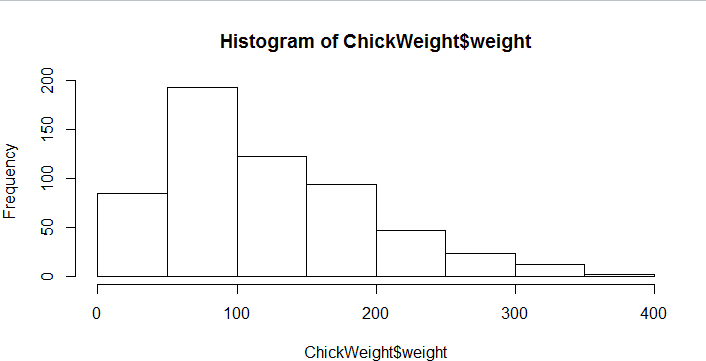
**SP and Weight(WT)**



**Sol:** Skewness for speed= -0.1139548, skewness value is negative so it is left skewed. Since magnitude is slightly greater than 0 it is slightly left skewed

And for distance= 0.7824835, right skewed (Positive) slight magnitude to right.

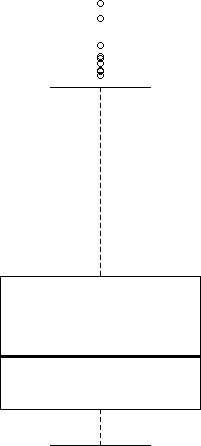
# Q10) Draw inferences about the following boxplot & histogram



**Sol:** The most of the data points are concerated in the range 50-100 with frequency 200. And least range of weight is 400 somewere around 0-10.

So the expected value the above distribution is 75.

Skewness- we can notice a long tail towards right so it is heavily right skewed.



Sol: Medican is less than mean right skewed and we have outlier on the upper side of box plot and there is less data points between Q1 and bottom point.

**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 ?

Sol: X+/-(Z1- α. σ/sqrt(n)

Degrees of freedom= 2000-1= 1999

Confidence interval= 94%

(1- σ/2)= 1-0.03) =0.97

for confidene interval for 94% is 1.882 Confidence interval for 98%= 2.33

Confidence interval for 96% = 2.05

**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?

Mean= 41, Median= 40, variance= 24.111, Standard deviation= 4.910

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

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

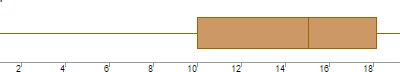
Q15) What is the nature of skewness when median > mean? Left Skewed

Q16) What does positive kurtosis value indicates for a data ? The data is notmally distributed and kurtosis value is 0.

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

The distribution of the data has lighter tails and a flatter peaks than the normal distribution.

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



What can we say about the distribution of the data?

Sol: Let’s assume above box plot is about age’s of the students in a school. 50% of the people are above 10 yrs old and remainig are less.

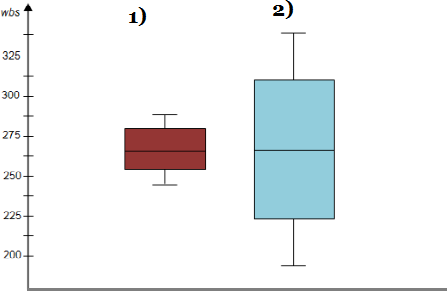
And students who’s age is above 15 are approx 40%.

What is nature of skewness of the data?

Sol: Left skewed, median is greater than mean.

What will be the IQR of the data (approximately)? Approximately= -8

Q19) Comment on the below Boxplot visualizations?



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

Sol: By observing both the plots whisker’s level is high in boxplot 2, mean and median are equal hence distribution is symetrical.

Q 20) Calculate probability from the given dataset for the below cases Data \_set: Cars.csv FILTERING

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)

Sol: By using filter command and installing the dplyr package into the ‘R’.

1. There are 33 observations in MPG which are greater than 38
2. 61 observations in MPG which are lesser than 40

c)

Q 21) Check whether the data follows normal distribution

* 1. Check whether the MPG of Cars follows Normal Distribution Dataset: Cars.csv
  2. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

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

Z score of 90% confidence interval is 1.65 Z score of 94% confidence interval is 1.55 Z score of 60% confidence interval is 0.85

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

For 95%= 1.96

For 96%= 2.5

For 99% = 2.47

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