**Excelr Data Science**

**Assignment - 1: Basic Statistics\_Level 1**

**Submitted by: Umadevi Betageri**

Q1) Identify the Data type for the following:

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

**-------------------------------------**

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 | Ordinal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ratio |
| Years of Education | Ordinal |

**-----------------------------------------**

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

Ans: Possible outcomes :

(H,H,H), (H,H,T), (H,T,H), (H, T, T), (T, H, H), (T, H, T), (T, T, H), (T, T, T)

There are 3 possibilities to get two heads and one tail out of eight possible combinations.

Hence, **P = 3/8**

**------------------------------------------**

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

All possible outcomes are

| (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6) |
| --- |
| (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6) |
| (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6) |
| (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6) |
| (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6) |
| (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) |

1. Probability that the sum is equal to 1

Ans: **P = 0**, Since the least sum of all possible combinations is 2.

1. Less than or equal to 4

Ans:There are a total 6 combinations of all possible outcomes to get the sum <= 4.

(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (3, 1)

Hence, **P = 6/36 = 1/6**

1. Sum is divisible by 2 and 3

Ans: Here 6 and 12 are the only two integers that are divided by both 2 and 3, since the maximum sum out of all possible combinations is 12. There are a total 6 combinations of all possible outcomes to get the sum divisible by 2 and 3.

Possible combinations to get a sum of 6 : (1, 5), (2, 4), (3, 3), (4, 2), (5, 1)

Possible combinations to get a sum of 12 : (6, 6)

Hence, P = 5/36 + 1/36 = 6/36

**P = 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?

Ans: The possible combinations of drawing two balls out of all 7 balls is 7C2

The possible combinations of drawing 2 balls, with none of them are blue is 5C2

Hence the probability that none of the balls drawn is blue is

P = 5C2 / 7C2 = 10/21

**P 10/21**

**------------------------------------------**

Q6) Calculate the Expected number of candies for a randomly selected child. Below are the probabilities of the count of candies for children(ignoring the nature of the child-Generalized view).

| CHILD | Candies count - P(Bi) | Probability - P(A|Bi) |
| --- | --- | --- |
| 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 is ∑ P(Bi) x P(A|Bi)

= 1 x 0.015 + 4 x 0.20 + 3 x 0.65 + 5 x 0.005 + 6 x 0.01 + 2 x 0.120

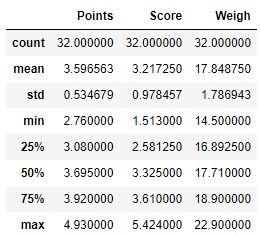
**Expected number of candies = 3.095**

**--------------------------------------------**

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

**Use Q7.csv file**

Ans: Mean, Median, Mode, Variance, Standard Deviation and Range

****

|  | **Points** | **Score** | **Weigh** |
| --- | --- | --- | --- |
| **Mean** | 3.596563 | 3.217250 | 17.848750 |
| **Median** | 3.6950 | 3.3250 | 17.710 |
| **Std. Deviation** | 0.534679 | 0.978457 | 1.786943 |
| **Variance** | 0.285881 | 0.957379 | 3.193166 |
| **Mode** | 3.07, 3.92 | 3.44 | 17.02, 18.90 |
| **Range** | 2.17 | 3.911 | 8.4 |

**Comments:**

1. Since the means of Points and Scores are slightly less than their median the distributions are left skewed. Whereas the mean of Weigh is almost equal to its median the distribution of Weigh is nearly normally distributed.
2. Points and Weigh have the 2 modes, whereas Score has only 1 mode.
3. Since the standard deviations of all variables are small, the data points are less dispersed.

**--------------------------------------------**

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: Expected value = Sum of all X (Weights) / n (Number of data points)

= (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199) / 9

**Expected value = 145.33 pounds**

**--------------------------------------------**

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

**Cars speed and distance. Use Q9\_a.csv**

**Ans:**

|  | **Speed** | **Distance** |
| --- | --- | --- |
| **Skewness** | -0.117510 | 0.806895 |
| **Kurtosis** | -0.508994 | 0.405053 |

**Comments:**

1. A skewness value of speed (-0.117) is slightly negative, which means that the data distribution is roughly symmetric or only slightly skewed to the left.
2. A skewness value of distance (0.81) is positive, indicating a right-skewed distribution. The tail of the distribution is longer on the right side, and the majority of the data points are concentrated on the left side.
3. For speed with a kurtosis of -0.51 indicates that the distribution of data is platykurtic. It has thinner tails and fewer outliers than a normal distribution
4. For distance with a kurtosis of 0.40 indicates that the distribution of data is leptokurtic. It has wider tails and more outliers than a normal distribution.

**SP and Weight(WT). Use Q9\_b.csv**

**Ans:**

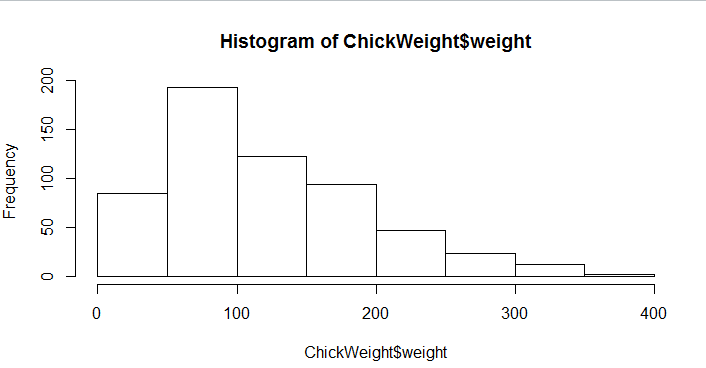
|  | **SP** | **Weight(WT)** |
| --- | --- | --- |
| **Skewness** | 1.611450 | -0.614753 |
| **Kurtosis** | 2.977329 | 0.950291 |

**Comments:**

1. A skewness value of 1.61 indicates that the distribution of SP data is positively skewed and the tail of the distribution is longer on the right side. The mean is typically greater than the median for this variable.
2. A skewness value of -0.61, indicates that the distribution of Weight (WT) data is negatively skewed and the tail of the distribution is longer on the left side. The mean is typically less than the median for this variable.
3. A kurtosis value of 2.97 indicates that the distribution of SP data is leptokurtic. In this case the tails are wider than those of a normal distribution. It means that there are more outliers.
4. A kurtosis value of 0.95 suggests that the distribution of Weight (WT) is platykurtic. In this case, the tails are lighter (less extreme) than those of a normal distribution and there are fewer outliers.

**--------------------------------------------**

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



**Ans: Comments**

1. The histogram of Weight indicates that the data is right skewed and has a wider tail on the right side.
2. The mean weight is greater than the median weight.
3. Maximum data points lie between 50 to 100 and very few data points lie between 300 to 400
4. This also indicates that there are some outliers.



**Ans: Comments**

1. The box plot indicates that there exists more than 7 outliers and the data is right skewed.

**--------------------------------------------**

**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: Here the standard deviation of the sample is given, hence we use t-distribution.

Confidence interval = 𝔁 ± t ( s / √n)

where, 𝔁 = sample mean = 200

t = t score

s = sample standard deviation = 30

n = sample size = 2000

degree of freedom = 200 - 1 = 199

1. t score for 94% confidence interval t = 1.8916 with 199 degrees of freedom.

hence, Confidence interval (94%) = 200 - 1.8916 x (30/√2000) = 198.731

200 + 1.8916 x (30/√2000) = 201.269

1. t score for 98% confidence interval t = 2.3452 with 199 degrees of freedom.

hence, Confidence interval (94%) = 200 - 2.3452 x (30/√2000) = 198.426

200 + 2.3452 x (30/√2000) = 201.573

1. t score for 98% confidence interval t = 2.0673 with 199 degrees of freedom.

hence, Confidence interval (94%) = 200 - 2.0673 x (30/√2000) = 198.622

200 + 2.0673 x (30/√2000) = 201.377

1. The **94%** confidence interval = (198.73, 201.27).
2. The **98%** confidence interval = (198.43, 201.57).
3. The **96%** confidence interval = (198.62, 201.38).

**--------------------------------------------**

**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: mean = 41.000000

median = 40.500000

variance = 25.529412

standard deviation = 5.052664

1. What can we say about the student marks?

Ans: Since the mean is very close to the median we can say that the student's score is nearly normally distributed.

**--------------------------------------------**

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

Ans: Data is Normally Distributed

**--------------------------------------------**

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

Ans: Right skewed or positively skewed

**--------------------------------------------**

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

Ans: Left skewed or negatively skewed

**--------------------------------------------**

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

Ans: Data has a steep peak.

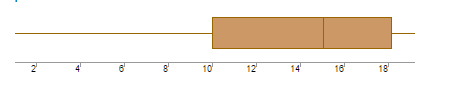
**--------------------------------------------**

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

Ans: The distribution is flat and has the lesser peak

**--------------------------------------------**

Q18) Answer the below questions using the below box plot visualization.



What can we say about the distribution of the data?

Ans: In this box plot even though there are no outliers, the data is not normally distributed. It is skewed towards the left and has a wider tail on the left. Hence the mean is less than the median.

What is the nature of skewness of the data?

Ans: Negative or left skewed

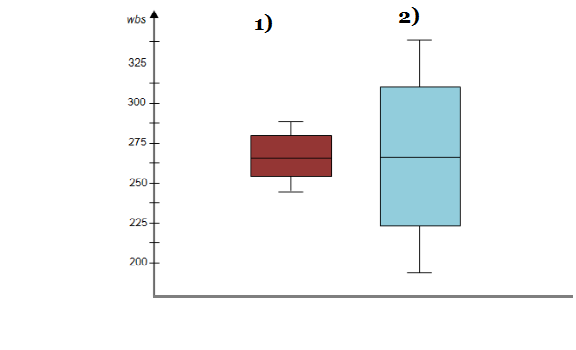
What will be the IQR of the data (approximately)?

Ans: IQR = Q3 - Q1

Approx. IQR = 18 - 10 = 8

**--------------------------------------------**

Q19) Comment on the below Box Plot visualizations?



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

Ans: Box plots of variables 1 and 2 show that they do not have any outliers and the data is normally distributed i.e., the mean and median are almost the same.

**--------------------------------------------**

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

**Ans : From the python coding we get the following details:**

**Mean of MPG = 34.422**

**Standard Deviation of MPG = 9.131**

To calculate this probability, let’s standardize the X values using the formula

Z = (X - μ) / σ,

where, X = 38,

μ = 34.422

σ = 9.131

* 1. P(MPG>38)

Ans: Z = (38 - 34.422) / 9.131

Z = 0.39185

From the standard normal distribution table probability for z = 0.39 is 0.6517

Hence,

P(MPG > 38) = 1 - 0.6517

**P(MPG > 38) = 0.348**

* 1. P(MPG<40)

Ans: Z = (40 - 34.422) / 9.131

Z = 0.611

From the standard normal distribution table probability for z = 0.611 is 0.7291

Hence,

**P(MPG < 40) = 0.729**

* 1. P (20<MPG<50)

Ans: Standardizing X values

**For X = 20**

Z = (20 - 34.422) / 9.131

Z = -1.5794

From the standard normal distribution table probability for z = -1.58 is 0.0571

**For X = 50**

Z = (50 - 34.422) / 9.131

Z = 1.7060

From the standard normal distribution table probability for z = 1.71 is 0.9564

Hence,

P(20 < MPG < 50) ≈ P(Z < 1.71) - P(Z < -1.58)

P(20 < MPG < 50) = 0.9564 - 0.0571

**P(20 < MPG < 50) = 0.8993**

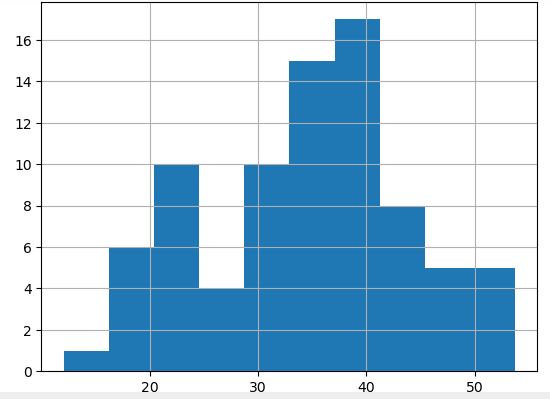
**--------------------------------------------**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

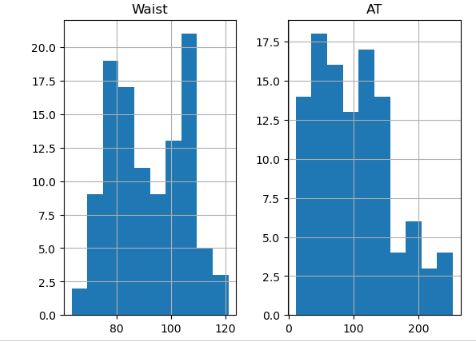
Ans: No, the MPG variable is not normally distributed, since the mean(34.42) is less than the median (35.152). The distribution has two peaks. The below plot shows the distribution of the MPG variable from cars dataset.



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

Dataset: wc-at.csv

Ans: No, both AT and Waist variables are not normally distributed. For both the variable mean is greater than median. Waist variable has the two peaks in the distribution, whereas the AT shows a right skewed distribution. The histogram plots of bo AT and Wait variables are shown below.



**--------------------------------------------**

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

Ans:

1. 90% confidence interval.

For a two tailed test, significance level α = 0.05

z value for α = 0.05 is 1.645 ( From Z table)

**For 90% confidence interval Z = 1.645**

1. 94 % confidence interval.

For a two tailed test, significance level α = 0.03

z value for α = 0.03 is 1.555 (From Z table).

**For 94% confidence interval Z = 1.555**

1. 60 % confidence interval.

For a two tailed test, significance level α = 0.2

z value for α = 0.2 is 0.253 (From Z table).

**For 60% confidence interval Z = 0.253**

**--------------------------------------------**

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

Ans: To find the t-score we need degrees of freedom (df).

df = 25 - 1 = 24. where 25 is the sample size.

1. 95% confidence interval.

From the t distribution table t score (two-tailed) with df = 24 is 2.064

**For 95% confidence interval t = 2.064**

1. 96% confidence interval.

From the t distribution table t score (two-tailed) with df = 24 is 2.171

**For 96% confidence interval t = 2.171**

1. 99% confidence interval.

From the t distribution table t score (two-tailed) with df = 24 is 2.797

**For 99% confidence interval t = 2.797.**

**--------------------------------------------**

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: Assume Null Hypothesis is: Ho = Avg. life of Bulb >= 260 days and Alternate Hypothesis is: Ha = Avg. life of Bulb < 260 days

We use following formula to find t score

**t =(x(bar) - μ)/(s / √n )**

Where, x(bar) = Sample mean

μ = Population mean

s = Sample standard deviation

n = sample size

t-scores at x = 260

t **=** (260 **-** 270) **/** (90 **/ (√**18))

t = -0.4714

Probability of that Avg. life of Bulb < 260 days

Using the t-distribution calculations form python code for degrees of freedom = 17.

**P(Avg. life of Bulb < 260) = 0.321**

**--------------------------------------------**