# Practice Problems

**Subject:** Statistics for Artificial Intelligence Data Science

**Subject Code:** CSDLO5011

| Course Code | Course Name | Credit |
|---|---|---|
| CSDLO5011 | Statistics for Artificial IntelligenceData Science | 03 |

| Prerequisite: C Programming | |
|---|---|
| **Course Objectives:** The course aims: | |
| 1 | To Perform exploratory analysis on the datasets |
| 2 | To Understand the various distribution and sampling |
| 3 | To Perform Hypothesis Testing on datasets |
| 4 | To Explore different techniques for Summarizing Data |
| 5 | To Perform The Analysis of Variance |
| 6 | To Explore Linear Least Squares |
| **Course Outcomes:** Learner will be able to | |
| 1 | Illustrate Exploratory Data Analysis |
| 2 | Describe Data and Sampling Distributions |
| 3 | Solve Statistical Experiments and Significance Testing |
| 4 | Demonstrate Summarizing Data |
| 5 | Interpret the Analysis of Variance |
| 6 | Use Linear Least Squares |

**Prerequisite:** Discrete Structures and Graph Theory

| Module | | Detailed Content | Hours |
|---|---|---|---|
| 1 | | **Exploratory Data Analysis** | 5 |
| | 1.1 | Elements of Structured Data ,Further Reading ,Rectangular Data ,Data Frames and Indexes ,Nonrectangular Data Structures , Estimates of Location ,Mean ,Median and Robust Estimates , Estimates of Variability,Standard Deviation and Related Estimates ,Estimates Based on Percentiles , Exploring the Data Distribution ,Percentiles and Boxplots ,Frequency Tables and Histograms ,Density Plots and Estimates. | |
| | 1.2 | Exploring Binary and Categorical Data , Mode ,Expected Value, Probability ,Correlation ,Scatterplots ,Exploring Two or More Variables ,Hexagonal Binning and Contours (Plotting Numeric Versus Numerical Data) ,Two Categorical Variables ,Categorical and Numeric Data ,Visualizing Multiple Variables. | |
| 2 | | **Data and Sampling Distributions** | 6 |
| | 2.1 | Random Sampling and Sample Bias ,Bias ,Random Selection ,Size Versus Quality,Sample Mean Versus Population Mean ,Selection Bias ,Regression to the Mean ,Sampling Distribution of a Statistic ,Central Limit Theorem ,Standard Error ,The Bootstrap ,Resampling Versus Bootstrapping . | |
| | 2.2 | Confidence Intervals ,Normal Distribution ,Standard Normal and QQ-Plots ,Long-Tailed Distributions ,Student's t-Distribution ,Binomial Distribution ,Chi-Square Distribution ,F-Distribution ,Poisson and Related Distributions ,Poisson Distributions ,Exponential Distribution ,Estimating the Failure Rate ,Weibull Distribution .<br><br>**Self Study :** Problems in distributions. | |
| 3 | | **Statistical Experiments and Significance Testing** | 8 |
| | 3.1 | A/B Testing ,Hypothesis Tests ,The Null Hypothesis ,Alternative Hypothesis ,One-Way Versus Two-Way Hypothesis Tests ,Resampling ,Permutation Test ,Example: Web Stickiness,Exhaustive and Bootstrap Permutation Tests ,Permutation Tests: The Bottom Line for Data Science ,Statistical Significance and p-Values ,p-Value ,Alpha ,Type 1 and | |

| | | | |
|---|---|---|---|
| | | Type 2 Errors | |
| | 3.2 | Data Science and p-Values , t-Tests ,Multiple Testing ,Degrees of Freedom ,ANOVA ,F-Statistic ,Two-Way ANOVA , Chi-Square Test ,Chi-Square Test: A Resampling Approach ,Chi-Square Test: Statistical Theory ,Fisher's Exact Test ,Relevance for Data Science ,Multi-Arm Bandit Algorithm ,Power and Sample Size ,Sample Size .<br><br>**Self Study :** Testing of Hypothesis using any statistical tool | |
| 4 | | **Summarizing Data** | 6 |
| | 4.1 | Methods Based on the Cumulative Distribution Function , The Empirical Cumulative Distribution Function ,The Survival Function ,Quantile-Quantile Plots , Histograms, Density Curves, and Stem-and-Leaf Plots , Measures of Location. | |
| | 4.2 | The Arithmetic Mean ,The Median , The Trimmed Mean , M Estimates , Comparison of Location Estimates ,Estimating Variability of Location Estimates by the Bootstrap , Measures of Dispersion , Boxplots , Exploring Relationships with Scatterplots .<br><br>**Self Study :** using any statistical tool perform data summarization | |
| 5 | | **The Analysis of Variance** | 6 |
| | 5.1 | The One-Way Layout, Normal Theory; the F Test ,The Problem of Multiple Comparisons , A Nonparametric Method—The Kruskal-Wallis Test ,The Two-Way Layout , Additive Parametrization , Normal Theory for the Two-Way Layout ,Randomized Block Designs , A Nonparametric Method—Friedman's Test . | |
| 6 | | **Linear Least Squares** | 8 |
| | 6.1 | Simple Linear Regression, Statistical Properties of the Estimated Slope and Intercept , Assessing the Fit , Correlation and Regression , The Matrix Approach to Linear Least Squares , Statistical Properties of Least Squares Estimates , Vector-Valued Random Variables , Mean and Covariance of Least Squares Estimates , Estimation of $\sigma 2$, Residuals and Standardized Residuals , Inference about $\beta$ , Multiple Linear Regression—An Example , Conditional Inference, Unconditional Inference, and the Bootstrap , Local Linear Smoothing .<br><br>**Self Study :**Create a Linear Regression model for a dataset and display the error measures, Chose a dataset with categorical data and apply linear regression model | |



| **Textbooks:** | |
|---|---|
| 1 | Bruce, Peter, and Andrew Bruce. Practical statistics for data scientists: 50 essential concepts. Reilly Media, 2017. |
| 2 | Mathematical Statistics and Data Analysis John A. Rice University of California, Berkeley,Thomson Higher Education |
| **References:** | |
| 1 | Dodge, Yadolah, ed. Statistical data analysis and inference. Elsevier, 2014. |
| 2 | Ismay, Chester, and Albert Y. Kim. Statistical Inference via Data Science: A Modern Dive into R and the Tidyverse. CRC Press, 2019. |
| 3 | Milton. J. S. and Arnold. J.C., "Introduction to Probability and Statistics", Tata McGraw Hill, 4th Edition, 2007. |
| 4 | Johnson. R.A. and Gupta. C.B., "Miller and Freund's Probability and Statistics for Engineers", Pearson Education, Asia, 7th Edition, 2007. |
| 5 | A. Chandrasekaran, G. Kavitha, "Probability, Statistics, Random Processes and Queuing Theory", Dhanam Publications, 2014. |

| **Assessment**: | |
|---|---|
| **Internal Assessment:** | |
| Assessment consists of two class tests of 20 marks each. The first-class test is to be conducted when approx. 40% syllabus is completed and second class test when additional40% syllabus is completed. Duration of each test shall be one hour. | |
| **End Semester Theory Examination:** | |
| 1 | Question paper will consist of 6 questions, each carrying 20 marks. |
| 2 | The students need to solve a total of 4 questions. |
| 3 | Question No.1 will be compulsory and based on the entire syllabus. |
| 4 | Remaining question (Q.2 to Q.6) will be selected from all the modules. |

| **Useful Links** | |
|---|---|
| 1 | https://www.edx.org/course/introduction-probability-science-mitx-6-041x-2 |
| 2 | https://www.coursera.org/learn/statistical-inference |
| 3 | https://www.datacamp.com/community/open-courses/statistical-inference-and-data-analysis |

*** Suggestion: Laboratory work based on the above syllabus can be incorporated as a mini project in CSM501: Mini-Project.**

# Index

# Practice Problems: Modue-1

**1.** The mean of 6, 8, x + 2, 10, 2x - 1, and 2 is 9. Find the value of x and also thevalue of the observation in the data. (9, 11, 17)

2. The runs scored in a cricket match by 11 players is as follows:7, 16, 121, 51, 101, 81, 1, 16, 9, 11, 16

Find the mean, mode, median of this data. (Mean = 39 1/11; Mode = 16; Median = 16)

3. The mean of the following distribution is 26. Find the value of p and also thevalue of the observation.

| xi | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| fi | 3 | 3 | p | 7 | p - 1 | 4 |

Also, find the mode and the given data(2, 1)

4. If a die is rolled, then find the variance and standard deviation of thepossibilities

| | Population | Sample |
|---|---|---|
| Variance | $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$ | $S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$ |
| Standard Deviation | $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}}$ | $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ |

(Variance is σ2 = 2.917, and Standard deviation = √2.917 = 1.708)

5. Find the standard deviation of the average temperatures recorded over a five-day period last winter: 18, 22, 19, 25, 12 (The mean = 19.2)

(Standard deviation for the temperatures recorded is 4.9; the variance is 23.7)

6. A survey of 36 students of a class was done to find out the mode of transport used by them while commuting to the school. The collected data is shown in the table given below. Represent the data in the form of a bar graph.

| Mode of Transport | Number of Students |
|---|---|
| Cycle | 16 |
| School Bus | 10 |
| WalkingCar | 4 |

7. Construct a frequency distribution table for the following weights (in gm) of 30 oranges using the equal class intervals, one of them is 40-45 (45 not included). The weights are: 31, 41, 46, 33, 44, 51, 56, 63, 71, 71, 62, 63, 54, 53, 51, 43, 36, 38, 54, 56, 66, 71, 74, 75, 46, 47, 59, 60, 61, 63.

**(a)** What is the class mark of the class intervals 50-55?

**(b)** What is the range of the above weights?

**(c)** How many class intervals are there?

**(d)** Which class interval has the lowest frequency?

ANS:

| C.I. | 30-35 | 35-40 | 40-45 | 45-50 | 50-55 | 55-60 | 60-65 | 65-70 | 70-75 | 75-80 |
|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 2 | 2 | 3 | 3 | 5 | 3 | 6 | 1 | 4 | 1 |

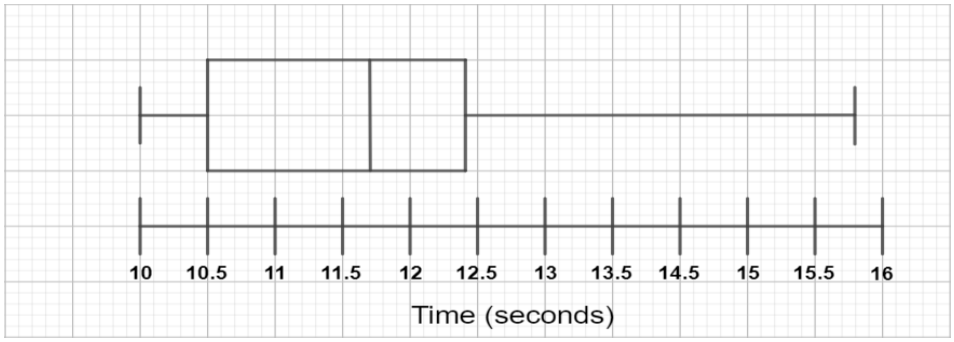
(a) 52.5

(b) 44 gm

(c) 10

(d) 65 - 70, 75 – 80

8.The box plot below was constructed from a collection of times taken to run a 100 m sprint. Using the box plot, determine the range and interquartile range.

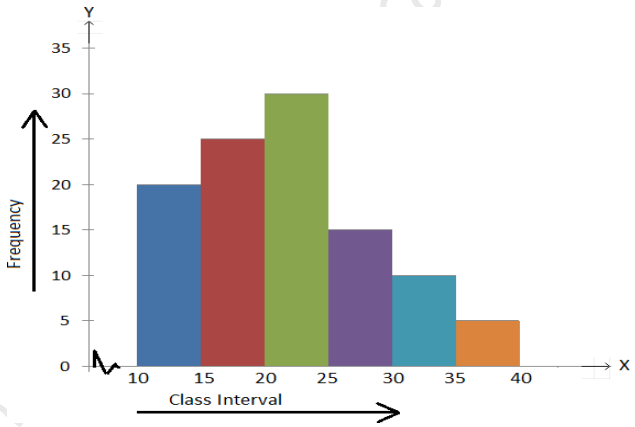



Ans :

Range =15.8−10=5.8 seconds.

Interquartile range =12.4−10.5=1.9 seconds

9. The histogram for a frequency distribution is given below




Answer the following.

**(i)** What is the frequency of the class interval 15 – 20?

**(ii)** What is the class intervals having the greatest frequency?

**(iii)** What is the cumulative frequency of the class interval 25 – 30?

**(iv)** Construct a short frequency table of the distribution.

**(v)** Construct a cumulative frequency table of the distribution

Solution:
(i) 25

(ii) 20 – 25

(iii) 90

10. In a certain property investment company with an international presence, workershave a mean hourly wage of $12 with a population standard deviation of $3. Given a sample size of 30, estimate and interpret the SE of the sample mean: (mean of $12 and a standard error of $0.55.)

11. Assume that we have increased the sample size to 80 in the example above and derived similar values for the mean and standard deviation of returns. Estimate the standard error of the sample mean.

A. 0.01

B. 0.02

C. 0.08

(The correct answer is A.)

12. X is a normally normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find

a) $P(x < 40)$

b) $P(x > 21)$

c) $P(30 < x < 35)$

a) 0.9938

b) 0.9878

c) 0.3944

13. A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr? (The probability that a car selected at a random has a speed greater than 100 km/hr isequal to 0.1587)

14. For a certain type of computers, the length of time bewteen charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. John owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.
(The probability that John's computer has a length of time between 50 and 70 hours isequal to 0.4082.)

15. Calculate the correlation coefficient for the following data. X = 4, 8 ,12, 16 and Y = 5, 10, 15
(Ans. 1)

16. Find the value of the correlation coefficient from the data given in thefollowing table:

| SUBJECT | AGE (x) | GLUCOSE LEVEL (y) |
|---|---|---|
| 1 | 43 | 99 |
| 2 | 21 | 65 |
| 3 | 25 | 79 |
| 4 | 42 | 75 |
| 5 | 57 | 87 |
| 6 | 59 | 81 |

(Ans-0.5298)

17. The scores for some candidates in a test are 40, 45, 49, 53, 61, 65, 71, 79, 85, 91.

What will be the percentile for the score 71?

(Ans-60)

18. The scores for some candidates in a test are 40, 45, 49, 53, 61, 65, 71, 79, 85, 91. What will be the score with a percentile value of 90?

(Ans-8)

# Practice Problems: Modue-2

**Central Limit Theorem**

**Bootstrap**

Confidence interval & Standard Error

1. **Find the standard error of the estimate of the mean weight of high school football players using the data given of weights of high school football players from your school. Then find a 95% confidence interval for the data.**

| Player Number | Weight in Pounds |
|---|---|
| 1 | 150 |
| 2 | 203 |
| 3 | 176 |
| 4 | 190 |
| 5 | 168 |
| 6 | 193 |
| 7 | 189 |
| 8 | 178 |
| 9 | 197 |
| 10 | 172 |

Ans.

Mean = 181.6 pounds, SD = 15.88 Standard error = 5.02 pounds

Confidence interval : We add & subtract 1.96 x 5.02.

Therefore it is 171.76 & 191.4

2. Find the standard error of the estimate for the average number of children in a household in your city by using the data collected from a sample of households in your city. Then find a 95% confidence interval for the data.

| Household | Number of Children |
|---|---|
| 1 | 2 |
| 2 | 3 |
| 3 | 1 |
| 4 | 0 |
| 5 | 5 |
| 6 | 2 |
| 7 | 1 |
| 8 | 4 |

Ans.

Mean = 2.23, SD = 1.669

Standard error = 0.59

Confidence interval : We add & subtract 1.96 x 0.59.

Therefore it is 1.09 & 3.4

## Normal Distribution & Standard Normal distribution

19. X is a normally distributed variable with mean $\mu = 30$ and standard deviation $\sigma = 4$.

Find a) $P(x < 40)$, b) $P(30 < x < 35)$

Ans:

a) 0.9938

b) 0.3944

20. A radar unit is used to measure speeds of cars on a motorway. The speeds are normally distributed with a mean of 90 km/hr and a standard deviation of 10 km/hr. What is the probability that a car picked at random is travelling at more than 100 km/hr?

Ans : (The probability that a car selected at a random has a speed greater than 100 km/hr is equal to 0.1587)

21. For a certain type of computers, the length of time between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. A student owns one of these computers and wants to know the probability that the length of time will be between 50 and 70 hours.

Ans: (The probability that John's computer has a length of time between 50 and 70 hours is equal to 0.4082.)

22. Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the GMAT?

Ans: 0.5948

23. How high must an individual score on the GMAT in order to score in the highest 5%?

Ans: 711.24

The length of human pregnancies from conception to birth approximates a normal distribution with a mean of 266 days and a standard deviation of 16 days. What proportion of all pregnancies will last between 240 and 270 days (roughly between 8 and 9 months)?

24.

Ans: 0.5471

25.

What length of time marks the shortest 70% of all pregnancies?

Ans:274.3226.

26.

The average number of acres burned by forest and range fires in a large New Mexico county is 4,300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2,500 and 4,200 acres will be burned in any given year?

Ans:0.440127.

27.

What number of burnt acres corresponds to the $38^{th}$ percentile?

Ans:4067.5

**T-distribution**

1. If the sample mean and expected mean value of the marks obtained by 15 studentsin a class test is 290 and 300 respectively. What is the t-score if the standard deviation of the marks is 50?

Answer: T score of the marks is -0.7745

2. If the sample mean and expected mean value of the marks obtained by 15 studentsin a class test is 290 and 300 respectively. What is the t-score if the standard deviation of the marks is 50?

Answer: T score of the marks is -0.7745.

3. If the sample mean and expected mean value of the height of 16 friends is 170 and 165 respectively. What is the t-score if the standard deviation of the heights is 21.05?

Answer: T score of the height is 0.95.

4. If the sample mean and expected mean value of the marks obtained by 15 studentsin a class test is 290 and 300 respectively. What is the t-score if the standard deviation of the marks is 50?

Answer: T score of the marks is -0.7745.

5. If the sample mean and expected mean value of the height of 16 friends is 170 and165 respectively. What is the t-score if the standard deviation of the heights is 21.05?

Answer: T score of the height is 0.95.

**QQ-Plots**

**Binomial distribution**

**Exponential distribution**

**A mobile conversation follows an exponential distribution $f(x) = \frac{1}{3}e^{-\frac{1}{a}x}$. What is the probability that the conversation takes more than 5 minutes?**

**Poisson distribution**

**F distribution**

**Chi square distribution**

**Weibull distribution**

6. Let X and Y be independent and identically distributed Poisson random variableswith rate λλ. Let T=X+YT=X+Y. Find the PMF of T.

7. In a sample of 8 observations, the entirety of squared deviations of things from themean was 94.5. In another specimen of 10 perceptions, the worth was observed to be 101.7 Test whether the distinction is huge at 5% level. (You are given that at 5% level of centrality, the basic estimation of FF for v1v1 = 7 and v2v2 = 9, F.05F.05 is3.29).

8. A poker-dealing machine is supposed to deal cards at random, as if from an infinite deck. In a test, you counted 1600 cards, and observed the following:

Spades 404

Hearts 420

Diamonds 400

Clubs 376

Could it be that the suits are equally likely? Or are these discrepancies too much tobe random?

9. Same as before, but this time jokers are included, and you counted 1662 cards, withthese results:

Spades 404

Hearts 420

Diamonds 400

Clubs 356

Jokers82

a. How many jokers would you expect out of 1662 random cards? How many ofeach suit?

b. Is it possible that the cards are really random? Or are the discrepancies toolarge?

10. A genetics engineer was attempting to cross a tiger and a cheetah. She predicted a phenotypic outcome of the traits she was observing to be in the following ratio 4 stripes only: 3 spots only: 9 both stripes and spots. When the cross was performed and she counted the individuals she found 50 with stripes only, 41 with spots only and 85 with both. According to the Chi-square test, did she get the predicted outcome?

11. Let X= amount of time a shopkeeper spends with his customer follows exponential distribution with the average amount of time equal to 4 minutes. Find theprobability that the shopkeeper is going to spend 5 minutes with the customer?

12. The amount of time a student takes to solve any problem follows an exponential distribution with the average amount of time equal to 8 minutes. What will be the probability that he will take 5 minutes to solve the problem?

13. Let $X$ be a random variable with mean $\mu$=20 and standard deviation $\sigma$=4. A sample of size 64 is randomly selected from this population. What is the approximateprobability that the sample mean $\bar{X}$ of the selected sample is less than 19?

14. In the first semester of the year 2003, the average return for a group of 251 investing companies was 4.5% and the standard deviation was 1.5%. If a sample of 40 companies is randomly selected from this group, what is the approximate probability that the average return of the companies in this sample wasbetween 4% and 5% in the first semester of the year 2003?

15. A pension fund company carries out a study of a large group of mutual funds and find that their average return over a period of 5 years was 80% 80% with a standarddeviation equal to 30%30%. If a sample of 5050 mutual funds is randomly selected from the group, what is the approximate probability that the sample had an averagereturn greater than 90% 90% over the 5 year period?

16. Assume that we have increased the sample size to 80 in the example above and derived similar values for the mean and standard deviation of returns. Estimate the standard error of thesample mean.

A. 0.01

B. 0.02

C. 0.08

(The correct answer is A.)

17 The Edwards's Theater chain has studied its movie customers to determine how much money they spend on concessions. The study revealed that the spending distribution is approximately normally distributed with a mean of $4.11 and a standard deviation of $1.37. What percentage of customers will spend less than $3.00 on concessions?

Ans:20.9%

# Module 3: Practice problems Part 1
# (Hypothesis Testing & Type I & 2 errors)

1. We have a medicine that is being manufactured and each pill is supposed tohave 14 milligrams of the active ingredient. What are our null and alternative hypotheses?
   *Ho $\mu$ = 14 mgHa $\mu \neq$ 14 mg*

2. The school principal wants to test if it is true what teachers say – that highschool juniors use the computer an average 3.2 hours a day. What are our null and alternative hypotheses?
   *Ho $\mu$ = 3.2 hrsHa $\mu \neq$ 3.2 hrs*

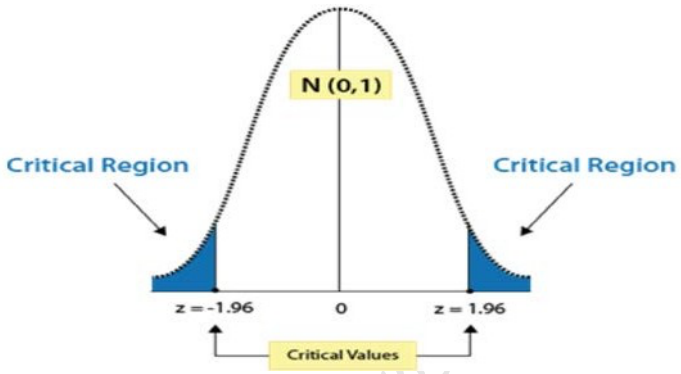
3. A researcher claims that black horses are, on average, more than 30 lbsheavier than white horses, which average 1100 lbs. What is the null hypothesis, and what kind of test is this?
   *The null hypothesis would be notated H0 : $\mu \leq$ 1130 lbs This is a right-tailed test, since the tail of the graph would be on the right. Recognize that values above 1130 would indicate that the null hypothesisbe rejected.*

4. A package of gum claims that the flavor lasts more than 39 minutes. What would be the null hypothesis of a test to determine the validity of the claim?What sort of test is this?
   *The null hypothesis would by notated as H0 : $\mu \leq$ 39. This is a right-tailed test, since the rejection region would consist of values greater than 39*

5. What is the critical value $\left(Z_{\frac{\alpha}{z}}\right)$ for a 95% confidence level, assuming a two- tailed test?

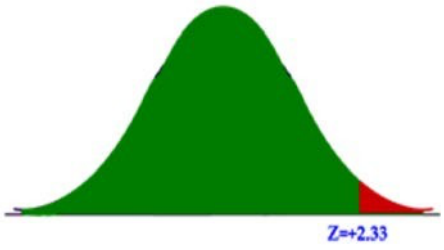
   A 95% confidence level means that a total of 5% of the area under the curve is considered the critical region. Since this is a two-tailed test, 1 2 of 5% = 2.5% of the values would be in the left tail, and the other 2.5% would be in the right tail. Looking up the Z-score associated with 0.025 on a reference table, we find 1.96. Therefore, +1.96 is the critical value of the right tail and -1.96 is the critical value of the left tail. The critical value for a 95% confidence level is Z = +/−1.96

6. Sketch the Z-score critical region for Example 5.




7. What would be the critical value for a right-tailed test with α = 0.01?

   *If α = 0.01, then the area under the curve representing H1, the alternative hypothesis, would be 99%, since α (alpha) is the same as the area of the rejection region. Using the Z-score reference table above, we find that the Z-score associated with 0.9900 is approximately 2.33. It appears that the critical value is Z = 2.33*



8. The school nurse thinks the average height of 7th graders has increased. The average height of a 7th grader five years ago was 145 cm with a standard deviation of 20 cm. She takes a random sample of 200 students and finds that the average height of her sample is 147 cm. Are 7th graders now taller than they were before? Conduct a single-tailed hypothesis test using a .05 significance level to evaluate the null and alternative hypotheses.

H0 : μ ≤ 145 Ha : μ > 145

*Choose α = .05. The critical value for this one tailed test is z=1.64. This is a one-tailed test, and a z-score of 1.64 cuts off 5% in the single tail. Any test statistic greater than 1.64 will be in the rejection region*

*Next, we calculate the test statistic for the sample of 7th graders. z = 147−145 √ 20 200 ≈ 1.414 The calculated z−score of 1.414 is smaller than 1.64 and thus does not fall in the critical region. Our decision is to fail to reject the null hypothesis and conclude that the probability of obtaining a sample mean equal to 147 is likely to have been due to chance.*

9. A farmer is trying out a planting technique that he hopes will increase the yield on his pea plants. The average number of pods on one of his pea plants is 145 pods with a standard deviation of 100 pods. This year, after trying his new planting technique, he takes a random sample of his plants and finds the average number of pods to be 147. He wonders whether or not this is a statistically significant increase. What are his hypotheses and the test statistic?

*H0 : μ ≤ 145 Ha : μ > 145*

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} = \frac{(147 - 145)}{\frac{100}{\sqrt{144}}} \approx 0.24$$

If we choose α = .05 4. The critical value will be 1.645. We will reject the null hypothesis if the test statistic is greater than 1.645. The value of the test statistic is 0.24.

This is less than 1.645 and so our decision is to fail to reject H0. Based on our sample we believe the mean is equal to 145.

10. The high school athletic director is asked if football players are doing as well academically as the other student athletes. We know from a previous study that the average GPA for the student athletes is 3.10. After an initiative to help improve the GPA of student athletes, the athletic director randomly samples 20 football players and finds that the average GPA of thesample is 3.18 with a sample standard deviation of 0.54. Is there a significant improvement? Use a 0.05 significance level.

H0 : μ = 3.10 Ha : μ 6= 3.10

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{3.18 - 3.10}{\frac{0.54}{\sqrt{20}}} = 0.66$$

We know that we have 20 observations, so our degrees of freedom for this test is 19. Nineteen degrees of freedom at the 0.05 significance level gives us a critical value of ± 2.093.

Thus, the athletic director can conclude that the mean academic performance of football players does not differ from the mean performance of other student athletes.

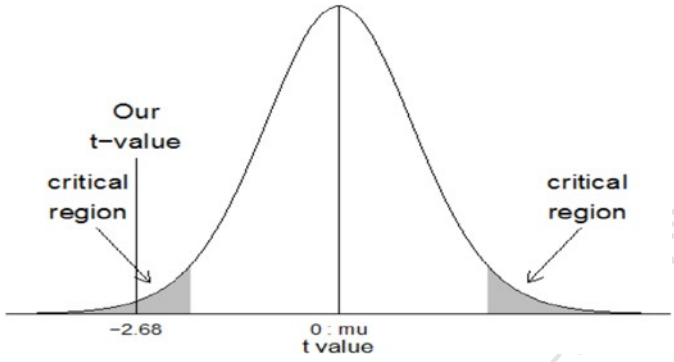
11. Duracell manufactures batteries that the CEO claims will last an average of 300 hours under normal use. A researcher randomly selected 20 batteries from the production line and tested these batteries. The tested batteries had a mean life span of 270 hours with a standard deviation of 50 hours. Do we have enough evidence to suggest that the claim of an average lifetime of 300 hours is false?

H0 : μ = 300 HA : μ 6= 300

Standard Error: $SEx^{-} = \sqrt{s}\ n\ SEx^{-} = \sqrt{}\ 50\ 20 = 11.18$

$t = x^{-} - \mu\ SEx^{-} = 270 - 300\ 11.18 = -2.68$

We know that we have 20 batteries, so our degrees of freedom for this test is (20-1)= 19. Nineteen degrees of freedom at the 0.05 significance level gives us a critical value of ± 2.093




The average battery life of the sample is significantly different from the average battery life claim by the CEO.

12. You have just taken ownership of a pizza shop. The previous owner told you that you would save money if you bought the mozzarella cheese in a 4.5 pound slab. Each time you purchase a slab of cheese, you weigh it to ensure that you are receiving 72 ounces of cheese. The results of 7 random measurements are 70, 69, 73, 68, 71, 69 and 71 ounces. Are these differences due to chance or is the distributor giving you less cheese than you deserve?
State the hypotheses.
Calculate the test statistic.
Would the null hypothesis be rejected at the 10% level? The 5% level? The 1% level?

a. H0 : $\mu = 72$; and for Ha : $\mu \neq 72$.

b. -2.9315.

c. The null hypothesis would be rejected at the .10 and the .05 levels, but not at the .01 level.

13. The average weight of a dumbbell in a gym is 90lbs. However, a physical trainer believes that the average weight might be higher. A random sample of 5 dumbbells with an average weight of 110lbs and a standard deviation of 18lbs. Using hypothesis testing check if the physical trainer's claim can be supported for a 95% confidence level.

The average weight of the dumbbells may be greater than 90lbs

14. The average score on a test is 80 with a standard deviation of 10. With a new teaching curriculum introduced it is believed that this score will change. On random testing, the score of 38 students, the mean was found to be 88. With a 0.05 significance level, is there any evidence to support this claim?

There is a difference in the scores after the new curriculum was introduced.

15. The average score of a class is 90. However, a teacher believes that the average score might be lower. The scores of 6 students were randomly measured. The mean was 82 with a standard deviation of 18. With a 0.05 significance level use hypothesis testing to check if this claim is true. There is not enough evidence to support the claim.

16. A stenographer claims that she can take dictation at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trials in which she demonstrated a mean of 116 words with standard deviation of 15 words ?

Claim rejected

17. An automatic machine was designed to pack exactly 2 kg. of tea. A sample of 100 packs was examined to test the machine. The average weight was found to be 1.94 kg. with standard deviation of 0.10 kg. is the machine working properly ?

The machine is not working properly

18. A sample of 600 persons selected at random from a large city shows that there are 53% smokers. Is there any reason to doubt the hypothesis that smokers and non-smokers are equal in number in the city ?

smokers and non-smokers are equal in numbers in that city

19. When flipped 1000 times, a coin landed 515 times heads up. Does it support the hypothesis that the coin is unbiased ?

The coin is not unbiased

20. While throwing 5 die 40 times, a person got success 25 times - getting a 4 was called success. Can we consider the difference between expected value and observed value as being significantly different ?

The dice is not unbiased

21. A patented medicine claimed that it is effective in curing 90% of the patients suffering from malaria. From a sample of 200 patients using this medicine, it was found that only 170 were cured. Determine whether the claim is right or wrong. (Take 1% level of significance).

The claim is justified

22. A random sample of 400 male students have average weight of 55 kg. Can we say that the sample comes from a population with mean 58 kg. with a variance of 9 kg. ?

The sample is not likely to be from the given population

23. A random sample of 400 tins of vegetable oil and labeled "5 kg. net weight" has a mean net weight of 4.98 kg. with standard deviation of 0.22 kg. Do we reject the hypothesis of net weight of 5 kg. per tin on the basis of this sample at 1% level of significance ?

Accepted at 1% level of significance

24. The maximum probability of committing a Type I error is

A. also the level of significance

B. never more than 0.05

C. the power of the test

D. zero if the null hypothesis is rejected

25. Which of the following is a correct statement (in the context of hypothesis tests)?

A. The Power of a test increases as the Type 2 error probability does

B. It is not possible to decrease both Type 1 error and Type 2 error at the same time.

C. The significance level is always equal to the probability of Type 2 error.

D. A test is significant if it fails to reject the null hypothesis.

26. Bottles of water have a label stating that the volume is 12 oz. A consumer group suspects the bottles are under-filled and plans to conduct a test. A Type I error in this situation would mean

A. the consumer group concludes the bottles have less than 12 oz. when the mean actually is 12 oz.

B. the consumer group does not conclude the bottles have less than 12 oz. when the mean actually is less than 12 oz.

C. the consumer group has evidence that the label is incorrect.

27. The owner of travel agency would like to determine whether or not the mean age of the agency's customers is over 24. If so, he plans to alter the destination of their special cruises and tours. If he concludes the mean age is over 24 when it is not, he makes a _____-error. If he concludes the mean age is not over 24 when it is, he makes a _____error.

A. Type II; Type II

B. Type I; Type I

C. Type I; Type II

D. Type II; Type I

28.Suppose we wish to test H : 53 vs H : 53 0 μ ≤ a μ > . What will result if we conclude that the mean is greater than 53 when its true value is really 55?

A. We have made a Type I error

B. We have made a correct decision

C. We have made a Type II error

D. None of the above are correct

29. A hypothesis test is used to prevent a machine from underfilling or overfilling quart bottles of beer. On the basis of sample, the machine is shut down for inspection. A thorough examination reveals there is nothing wrong with the filling machine. From a statistical point of view:

- Both Type I and Type II errors were made.
- A Type I error was made.
- A Type II error was made.
- A correct decision was made.

30. A bottling company needs to produce bottles that will hold 12 ounces of liquid. Periodically, the company gets complaints that their bottles are not holding enough liquid. To test this claim, the bottling company randomly samples 36 bottles. Suppose the p-value of this test turned out to be 0.0455. State the proper conclusion.

- At $\alpha = 0.085$, fail to reject the null hypothesis.
- At $\alpha = 0.035$, accept the null hypothesis.
- At $\alpha = 0.05$, reject the null hypothesis.
- At $\alpha = 0.025$, reject the null hypothesis.

31.Which of the following are A/B Testing tools?

- Visual Website optimizer
- Google Content Experiments
- Optimizely
- All of the above

32. Always perform A/B Testing if there is probability to beat the original variation by?

- 0.05
- less than 5%
- greater than 5%
- greater than equal to 5%

33. A weight reducing program that includes a strict diet and exercise claims on its online advertisement that it can help an average overweight person lose 10 pounds in three months. Following the program's method a group of twelve overweight persons have lost 8.11 5.7, 11.6, 12.9, 3.8, 5.9, 7.8, 9.1, 7.0, 8.2, 9.3 and 8.0 pounds in three months. Test at

5% level of significance whether the program's advertisement is overstating the reality.

Program is overstating the reality

34. A ketchup manufacturer is in the process of deciding whether to produce an extra spicy brand. The company's marketing research department used a national telephone survey of 6000 households and found the extra spicy ketchup would be purchased by 335 of them. A much more extensive study made two years ago showed that 5% of the households would purchase the brand then. At a 2% significance level, should the company conclude that there is an increased interest in the extra-spicy flavour?

Current interest is significantly greater than the interest 2 years ago.

35. A sample of 32 money market mutual funds was chosen on January 1, 1996 and the average annual rate of return over the past 30 days was found to be 3.23% and the sample standard deviation was 0.51%. A year earlier a sample of 38 money-market funds showed an average rate of return of 4.36%. Is it reasonable to conclude (at $\alpha = 0.05$) that money-market interest rates declined during 1995?

Reject Ho

36. A large hotel chain in trying to decide whether to convert more of its rooms into non-smoking rooms. In a random sample of 400 guests last year, 166 had requested the non-smoking rooms. This year 205 guests in a sample of 380 preferred the non-smoking rooms. Would you recommend that the hotel chain convert more rooms to non-smoking? Support your recommendation by testing the appropriate hypotheses at 0.01 level of significance.

Convert more rooms to Non-smoking.

# Module 3 part 2 : Practice problems on Anova

1. A clinical trial is run to compare weight loss programs and participants are randomly assigned to one of the comparison programs and are counselled on the details of the assigned program. Participants follow the assigned program for 8 weeks. The outcome of interest is weight loss, defined as the difference in weight measured at the start of the study (baseline) and weight measured at the end of the study (8 weeks), measured in pounds.

| Low Calorie | Low Fat | Low Carbohydrate | Control |
|---|---|---|---|
| 8 | 2 | 3 | 2 |
| 9 | 4 | 5 | 2 |
| 6 | 3 | 4 | -1 |
| 7 | 5 | 2 | 0 |
| 3 | 1 | 3 | 3 |

**ANSWER:**

*We reject $H_0$ because $8.43 \geq 3.24$. We have statistically significant evidence at $\alpha=0.05$ to show that there is a difference in mean weight loss among the four diets.*

2. Calcium is an essential mineral that regulates the heart, is important for blood clotting and for building healthy bones. The National Osteoporosis Foundation recommends a daily calcium intake of 1000-1200 mg/day for adult men and women. While calcium is contained in some foods, most adults do not get enough calcium in their diets and take supplements. Unfortunately some of the supplements have side effects such as gastric distress, making them difficult for some patients to take on a regular basis.

**A study is designed to test whether there is a difference in mean** daily calcium intake in adults with normal bone density, adults with osteopenia (a low bone density which may lead to osteoporosis) and adults with osteoporosis. Adults 60 years of age with normal bone density, osteopenia and osteoporosis are selected at random from hospital records and invited to participate in the study. Each participant's daily calcium intake is measured based on reported food intake and supplements. The data are shown below

| Normal Bone Density | Osteopenia | Osteoporosis |
|---|---|---|
| 1200 | 1000 | 890 |
| 1000 | 1100 | 650 |
| 980 | 700 | 1100 |
| 900 | 800 | 900 |
| 750 | 500 | 400 |
| 800 | 700 | 350 |

**ANSWER:**

*We do not reject $H_0$ because 1.395 < 3.68. We do not have statistically significant evidence at a =0.05 to show that there is a difference in mean calcium intake in patients with normal bone density as compared to osteopenia and osterporosis.*

3. Solve using One-way ANOVA method

| Observation | A | B | C | D |
|---|---|---|---|---|
| 1 | 8 | 12 | 18 | 13 |
| 2 | 10 | 11 | 12 | 9 |
| 3 | 12 | 9 | 16 | 12 |

| | | | | |
|---|---|---|---|---|
| 4 | 8 | 14 | 6 | 16 |
| 5 | 7 | 4 | 8 | 15 |

**ANSWER:**

*As calculated F=1.2821<3.2389*

*So, H0 is accepted, Hence there is no significant differentiating between samples*

4. Solve using One-way ANOVA method

| Observation | A | B | C |
|---|---|---|---|
| 1 | 8 | 7 | 6 |
| 2 | 10 | 7 | 8 |
| 3 | 6 | 8 | 10 |
| 4 | 7 | 9 | 6 |
| 5 | 9 | 8 | 4 |
| 6 | 0 | 5 | 5 |
| 7 | 0 | 0 | 7 |

**ANSWER:**

*As calculated F=1.0564<3.6823*

*So, H0 is accepted, Hence there is no significant differentiating between samples*

5. Solve using One-way ANOVA method

| Observation | A | B | C |
|---|---|---|---|
| 1 | 25 | 31 | 24 |
| 2 | 30 | 39 | 30 |
| 3 | 36 | 38 | 28 |
| 4 | 38 | 42 | 25 |
| 5 | 31 | 35 | 28 |

**ANSWER:**

*As calculated F=7.5>3.8853*

*So, H0 is rejected, Hence there is significant differentiating between samples*

6. Do ONE WAY ANOVA

Twenty-one students at the Autonomous University of Madrid (AUM) in Spain were selected for an informal study about student study skills; 7 first year, 7 second year, and 7 third year undergraduates were randomly selected.

The students were given a study-skills assessment having a maximum score of 100. As researches we are interested in whether or not a difference exists somewhere between the three different year levels. We will conduct this analysis using a One-Way ANOVA technique.

| col 1 | col 2 | col 3 |
|---|---|---|
| 82 | 71 | 64 |
| 93 | 62 | 73 |
| 61 | 85 | 87 |
| 74 | 94 | 91 |
| 69 | 78 | 56 |
| 70 | 66 | 78 |
| 53 | 71 | 87 |

**ANSWER:**

| **F** | 0.284805 |
|---|---|

7. Do TWO WAY ANOVA

Let's assume that Starbucks uses "secret shoppers" who appear to be customers to enter a store and document their experience in terms of customer service, cleanliness, and quality. The secret shoppers receive standardized training by Starbucks to ensure consistency and objectivity in their store reviews.

For its locations in the Australian cities of Sydney, Brisbane, and Melbourne, Starbucks has trained 6 secret shoppers. Each of the 6 secret shoppers will be assigned to visit the same store in each of the 3 cities. The visit sequence will be assigned randomly (hence *randomized* block design).

We would like to know if a **difference** in secret shopper ratings exists **among the cities**. Are they all about the same? Is one significantly higher than the other two? Are all three different from each other?

| | col 1 | col 2 | col 3 |
|---|---|---|---|
| Block-1 | 75 | 75 | 90 |
| Block-2 | 70 | 70 | 70 |
| Block-3 | 50 | 55 | 75 |
| Block-4 | 65 | 60 | 85 |
| Block-5 | 80 | 65 | 80 |
| Block-6 | 65 | 65 | 65 |

**ANSWER:**

| | |
|---|---|
| **F (MSC/MSE)** | 5.526316 |
| **F (MSB/MSE)** | 3.157895 |

F(MSC/MSE) critical value 4.1, hence null hypothesis is rejected

8. Explain briefly why use ANOVA?

9. What is the difference between one way & two way ANOVA test?

10. Write a short note on hypothesis testing.

11. What is Fisher's exact test & when is it used?

# Module 4 : Practice Problems

*(All problems to be solved by writing a python code as well exploring the same in Microsoft Excel)*

For data , famous  Gettysburg Address by Abraham Lincoln

Is given below :

**Gettysburg Address by Abraham Lincoln**

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battlefield of that war.  We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate—we cannot consecrate—we cannot hallow—this ground.  The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract.  The world will little note, nor long remember what we say here, but it can never forget what they did here.  It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us—that from these honoured dead we take increased devotion to that cause for which they gave the last full measure of devotion—that we here highly resolve that these dead shall not have died in vain—that this nation, under God, shall have a new birth of freedom—and that government of the people, by the people, for the people, shall not perish from the earth.

**Passage Data is given below :**

| Sr. no. of word | Letters in word |
|---|---|
| 1 | 4 |
| 2 | 5 |
| 3 | 3 |
| 4 | 5 |
| 5 | 5 |
| 6 | 3 |
| 7 | 3 |
| 8 | 7 |
| 9 | 7 |
| 10 | 5 |
| 11 | 4 |
| 12 | 4 |
| 13 | 9 |
| 14 | 1 |
| 15 | 3 |
| 16 | 6 |
| 17 | 9 |
| 18 | 2 |
| 19 | 7 |
| 20 | 3 |
| 21 | 9 |
| 22 | 2 |
| 23 | 3 |
| 24 | 11 |

| Sr. no. of word | Letters in word |
|---|---|
| 44 | 2 |
| 45 | 3 |
| 46 | 6 |
| 47 | 2 |
| 48 | 9 |
| 49 | 3 |
| 50 | 2 |
| 51 | 9 |
| 52 | 3 |
| 53 | 4 |
| 54 | 6 |
| 55 | 2 |
| 56 | 3 |
| 57 | 3 |
| 58 | 2 |
| 59 | 1 |
| 60 | 5 |
| 61 | 11 |
| 62 | 2 |
| 63 | 4 |
| 64 | 3 |
| 65 | 2 |
| 66 | 4 |
| 67 | 4 |

| Sr. no. of word | Letters in word |
|---|---|
| 87 | 4 |
| 88 | 4 |
| 89 | 6 |
| 90 | 5 |
| 91 | 4 |
| 92 | 2 |
| 93 | 2 |
| 94 | 10 |
| 95 | 7 |
| 96 | 3 |
| 97 | 6 |
| 98 | 4 |
| 99 | 2 |
| 100 | 5 |
| 101 | 2 |
| 102 | 4 |
| 103 | 3 |
| 104 | 2 |
| 105 | 1 |
| 106 | 6 |
| 107 | 5 |
| 108 | 2 |
| 109 | 6 |
| 110 | 8 |

| Sr. no. of word | Letters in word |
|---|---|
| 130 | 2 |
| 131 | 3 |
| 132 | 5 |
| 133 | 3 |
| 134 | 4 |
| 135 | 5 |
| 136 | 2 |
| 137 | 3 |
| 138 | 2 |
| 139 | 7 |
| 140 | 3 |
| 141 | 5 |
| 142 | 4 |
| 143 | 6 |
| 144 | 4 |
| 145 | 3 |
| 146 | 4 |
| 147 | 8 |
| 148 | 4 |
| 149 | 2 |
| 150 | 3 |
| 151 | 4 |
| 152 | 3 |
| 153 | 2 |

| | |
|---|---|
| 25 | 4 |
| 26 | 3 |
| 27 | 3 |
| 28 | 3 |
| 29 | 7 |
| 30 | 5 |
| 31 | 3 |
| 32 | 2 |
| 33 | 3 |
| 34 | 7 |
| 35 | 2 |
| 36 | 1 |
| 37 | 5 |
| 38 | 5 |
| 39 | 3 |
| 40 | 7 |
| 41 | 7 |
| 42 | 4 |
| 43 | 6 |

| | |
|---|---|
| 68 | 2 |
| 69 | 8 |
| 70 | 1 |
| 71 | 7 |
| 72 | 2 |
| 73 | 4 |
| 74 | 5 |
| 75 | 2 |
| 76 | 1 |
| 77 | 5 |
| 78 | 7 |
| 79 | 5 |
| 80 | 3 |
| 81 | 5 |
| 82 | 3 |
| 83 | 4 |
| 84 | 4 |
| 85 | 5 |
| 86 | 5 |

| | |
|---|---|
| 111 | 2 |
| 112 | 6 |
| 113 | 10 |
| 114 | 2 |
| 115 | 6 |
| 116 | 6 |
| 117 | 4 |
| 118 | 6 |
| 119 | 3 |
| 120 | 5 |
| 121 | 3 |
| 122 | 6 |
| 123 | 3 |
| 124 | 4 |
| 125 | 3 |
| 126 | 9 |
| 127 | 4 |
| 128 | 4 |
| 129 | 11 |

| | |
|---|---|
| 154 | 3 |
| 155 | 5 |
| 156 | 6 |
| 157 | 4 |
| 158 | 4 |
| 159 | 3 |
| 160 | 4 |
| 161 | 2 |
| 162 | 2 |
| 163 | 3 |
| 164 | 2 |
| 165 | 3 |
| 166 | 6 |
| 167 | 6 |
| 168 | 2 |
| 169 | 2 |
| 170 | 9 |
| 171 | 4 |
| 172 | 2 |

| Sr. no. of word | Letters in word |
|---|---|
| 173 | 3 |
| 174 | 10 |
| 175 | 4 |
| 176 | 5 |
| 177 | 4 |
| 178 | 3 |
| 179 | 6 |
| 180 | 4 |
| 181 | 4 |
| 182 | 4 |
| 183 | 3 |
| 184 | 2 |
| 185 | 5 |
| 186 | 8 |
| 187 | 2 |
| 188 | 2 |
| 189 | 6 |
| 190 | 3 |
| 191 | 2 |
| 192 | 2 |
| 193 | 2 |
| 194 | 4 |
| 195 | 9 |

| Sr. no. of word | Letters in word |
|---|---|
| 216 | 5 |
| 217 | 4 |
| 218 | 4 |
| 219 | 3 |
| 220 | 4 |
| 221 | 4 |
| 222 | 7 |
| 223 | 2 |
| 224 | 8 |
| 225 | 4 |
| 226 | 2 |
| 227 | 4 |
| 228 | 6 |
| 229 | 7 |
| 230 | 4 |
| 231 | 5 |
| 232 | 4 |
| 233 | 5 |
| 234 | 3 |
| 235 | 4 |
| 236 | 4 |
| 237 | 2 |
| 238 | 4 |

| Sr. no. of word | Letters in word |
|---|---|
| 259 | 6 |
| 260 | 3 |
| 261 | 3 |
| 262 | 6 |
| 263 | 5 |
| 264 | 3 |
| 265 | 6 |
| 266 | 4 |
| 267 | 3 |
| 268 | 5 |
| | |
| | |
| | |
| | |
| | |

| | | | | |
|---|---|---|---|---|
| 196 | 2 | | 239 | 4 |
| 197 | 3 | | 240 | 4 |
| 198 | 5 | | 241 | 6 |
| 199 | 4 | | 242 | 5 |
| 200 | 9 | | 243 | 3 |
| 201 | 6 | | 244 | 5 |
| 202 | 2 | | 245 | 4 |
| 203 | 4 | | 246 | 1 |
| 204 | 4 | | 247 | 3 |
| 205 | 5 | | 248 | 5 |
| 206 | 7 | | 249 | 2 |
| 207 | 4 | | 250 | 7 |
| 208 | 2 | | 251 | 3 |
| 209 | 4 | | 252 | 4 |
| 210 | 9 | | 253 | 10 |
| 211 | 8 | | 254 | 2 |
| 212 | 2 | | 255 | 3 |
| 213 | 4 | | 256 | 6 |
| 214 | 5 | | 257 | 2 |
| 215 | 3 | | 258 | 3 |



**USE ABOVE DATA TO SOLVE THE PROBLEMS GIVEN BELOW :**

Q.1 Find Mean, Mode, Median, Variance, Standard Deviation of the above population.

Q. 2 Find $10^{th}$, $25^{th}$, $50^{th}$, $75^{th}$ , $90^{th}$ percentile for the above data.

Q. 3 Plot Bar chart & Histogram for the above population

Q. 4 Plot Scattered Plot for above population. Find correlation coefficient between col.1 & Col. 2

Q. 5 Draw box plot for above population.

Q 6 Prints word numbers whose

a. Letters are less than or equal to 4
b. Letters are less than or equal to 10

Q 7 Calculate Z score for [4,5,6,6,6,7,8,12,13,13,14,18]

Q8. Draw scattered plot & find correlation coefficient for the following data :

| **x** | **y** |
|---|---|
| 14.2 | 215 |
| 16.4 | 325 |
| 11.9 | 185 |
| 15.2 | 332 |
| 18.5 | 406 |
| 22.1 | 522 |
| 19.4 | 412 |
| 25.1 | 614 |
| 23.4 | 544 |
| 18.1 | 421 |

Q9. A clinical trial is run to compare weight loss programs and participants are randomly assigned to one of the comparison programs and are counselled on the details of the assigned program. Participants follow the assigned program for 8 weeks. The outcome of interest is weight loss, defined as the difference in weight measured at the start of the study (baseline) and weight measured at the end of the study (8 weeks), measured in pounds. (one way Anova)

| **Low Calorie** | **Low Fat** | **Low Carbohydrate** | **Control** |
|---|---|---|---|
| 8 | 2 | 3 | 2 |

| | | | |
|---|---|---|---|
| 9 | 4 | 5 | 2 |
| 6 | 3 | 4 | -1 |
| 7 | 5 | 2 | 0 |
| 3 | 1 | 3 | 3 |

**ANSWER:**

*We reject $H_0$ because 8.43 $\geq$ 3.24. We have statistically significant evidence at $\alpha$=0.05 to show that there is a difference in mean weight loss among the four diets.*

**10.**Solve using One-way ANOVA

| Observation | A | B | C |
|---|---|---|---|
| 1 | 8 | 7 | 6 |
| 2 | 10 | 7 | 8 |
| 3 | 6 | 8 | 10 |
| 4 | 7 | 9 | 6 |
| 5 | 9 | 8 | 4 |
| 6 | 0 | 5 | 5 |
| 7 | 0 | 0 | 7 |

**Q 11.**

Let's assume that Starbucks uses "secret shoppers" who appear to be customers to enter a store and document their experience in terms of customer service, cleanliness, and quality. The secret shoppers receive standardized training by Starbucks to ensure consistency and objectivity in their store reviews.

For its locations in the Australian cities of Sydney, Brisbane, and Melbourne, Starbucks has trained 6 secret shoppers. Each of the 6 secret shoppers will be assigned to visit the same store in each of the 3 cities. The visit sequence will be assigned randomly (hence *randomized* block design).

We would like to know if a **difference** in secret shopper ratings exists **among the cities**. Are they all about the same? Is one significantly higher than the other two? Are all three different from each other?

| | col 1 | col 2 | col 3 |
|---|---|---|---|
| Block-1 | 75 | 75 | 90 |
| Block-2 | 70 | 70 | 70 |
| Block-3 | 50 | 55 | 75 |
| Block-4 | 65 | 60 | 85 |
| Block-5 | 80 | 65 | 80 |
| Block-6 | 65 | 65 | 65 |



**ANSWER:**

| | |
|---|---|
| **F (MSC/MSE)** | 5.526316 |
| **F (MSB/MSE)** | 3.157895 |

F(MSC/MSE) critical value 4.1, hence null hypothesis is rejected

# Module 5 : Practice Problems

## F-Test

1. Perform an F Test for the following samples.

i. Sample 1 with variance equal to 109.63 and sample size equal to 41.

ii. Sample 2 with variance equal to 65.99 and sample size equal to 21.

***Ans:*** *It is clear from the values that 1.66 < 2.287. Hence, the null hypothesis cannot be rejected.*

*2.* A research team wants to study the effects of a new drug on insomnia. 8 tests were conducted with a variance of 600 initially. After 7 months 6 tests were conducted with a variance of 400. At a significance level of 0.05 was there any improvement in the results after 7 months?

***Answer:*** *Fail to reject the null hypothesis.*

*3.* Pizza delivery times of two cities are given below
City 1: Number of delivery times observed = 28, Variance = 38
City 2: Number of delivery times observed = 25, Variance = 83
Check if the delivery times of city 1 are lesser than city 2 at a 0.05 alpha level.

***Answer:*** *Reject the null hypothesis.*

4. A toy manufacturer wants to get batteries for toys. A team collected 41 samples from supplier A and the variance was 110 hours. The team also collected 21 samples from supplier B with a variance of 65 hours. At a 0.05 alpha level determine if there is a difference in the variances.

***Answer:*** *Fail to reject the null hypothesis*

5. Let's say we have two data sets A & B which contains different data points. Perform F-Test to determine whether we can reject the null hypothesis at a 1% level of significance.

| | A | B |
|---|---|---|
| 3 | | |
| 4 | A | B |
| 5 | 93 | 12 |
| 6 | 50 | 11 |
| 7 | 53 | 20 |
| 8 | 92 | 31 |
| 9 | 21 | 65 |
| 10 | 1 | 10 |
| 11 | 2 | 3 |
| 12 | 85 | 9 |
| 13 | 86 | 1 |
| 14 | 22 | 4 |
| 15 | | 12 |
| 16 | | 87 |
| 17 | | 43 |
| 18 | | 23 |
| 19 | | 52 |
| 20 | | 49 |
| 21 | | 17 |
| 22 | | 17 |
| 23 | | 14 |
| 24 | | 24 |
| 25 | | |
| 26 | Level of significance (alpha) | 0.01 |
| 27 | | |

*Ans: So F critical value = 3.5225. Since F critical is greater than the F value, we cannot reject the null hypothesis.*

6. Suppose that you are working in a research company and want to the level of carbon oxide emission happening from 2 different brands of cigarettes and whether they are significantly different or not. In your analysis, you have collected the following information:

| | A | B | C |
|---|---|---|---|
| 4 | | | |
| 5 | | XYZ | ABC |
| 6 | Sample Size | 11 | 10 |
| 7 | Mean | 16.4 | 15.6 |
| 8 | Standard Deviation | 1.2 | 1.1 |
| 9 | | | |
| 10 | Level of Significance (alpha) | 0.05 | |
| 11 | | | |

*Ans: F Critical Value = 3.137. Since the F critical > F value, the null hypothesis cannot be rejected.*

**7.** A statistician was carrying out F-Test. He got the F statistic as 2.38. The degrees of freedom obtained by him were 8 and 3. Find out the F value from the F Table and determine whether we can reject the null hypothesis at 5% level of significance (one-tailed test).

*Ans: The F critical value obtained from the table is **8.845**. Since the F statistic (2.38) is lesser than the F Table Value (8.845), we cannot reject the null hypothesis.*

8. The bank has a Head Office in Delhi and a branch at Mumbai. There are long customer queues at one office, while customer queues are short at the other office. The Operations Manager of the bank wonders if the customers at one branch are more variable than the number of customers at another branch. A research study of customers is carried out by him.

The variance of Delhi Head Office customers is 31, and that for the Mumbai branch is 20. The sample size for Delhi Head Office is 11, and that for the Mumbai branch is 21. Carry out a two-tailed F-test with a level of significance of 10%.

*Ans: F critical value = 3.5225*

*Since F critical is greater than the F value, we cannot reject the null hypothesis.*

9. Two random samples were drawn from two normal populations ant their valure are given below. Test whether the two population have the same variance at 5% level of significance.

| A | B |
|---|---|
| 16 | 14 |
| 17 | 16 |
| 25 | 24 |
| 26 | 28 |
| 32 | 32 |
| 34 | 35 |
| 38 | 37 |
| 40 | 42 |
| 42 | 43 |
| | 45 |
| | 47 |



**Kruskal-Wallis test**

10. In a manufacturing unit, four teams of operators were randomly selected and sent to four different facilities for machining techniques training. After the training, the supervisor conducted the exam and recorded the test scores. At 95% confidence level does the scores are same in all four facilities?

| Facility 1 | Facility 2 | Facility 3 | Facility 4 |
|---|---|---|---|
| 88 | 77 | 71 | 52 |
| 82 | 76 | 56 | 65 |
| 86 | 84 | 64 | 68 |
| 87 | 59 | 51 | 81 |

*Ans: Calculated $\chi^2$ value is greater than the critical value of $\chi^2$ for a 0.05 significance level. $\chi^2_{calculated} > \chi^2_{critical}$ hence reject the null hypotheses.*

11. A researcher wants to know whether or not three drugs have different effects on knee pain, so he recruits 30 individuals who all experience similar knee pain and randomly splits them up into three groups to receive either Drug 1, Drug 2, or Drug 3.

After one month of taking the drug, the researcher asks each individual to rate their knee pain on a scale of 1 to 100, with 100 indicating the most severe pain.

The ratings for all 30 individuals are shown below:

| **Drug 1** | **Drug 2** | **Drug 3** |
|---|---|---|
| 78 | 71 | 57 |
| 65 | 66 | 88 |
| 63 | 56 | 58 |
| 44 | 40 | 78 |
| 50 | 55 | 65 |
| 78 | 31 | 61 |
| 70 | 45 | 62 |
| 61 | 66 | 44 |
| 50 | 47 | 48 |
| 44 | 42 | 77 |



*Ans: Since the p-value of the test (**0.21342**) is not less than 0.05, we fail to reject the null hypothesis.*

*We do not have sufficient evidence to say that there is a statistically significant difference between the median knee pain ratings across these three groups.*

12. We will use data on antibody production after receiving a vaccine. A hospital administered three different vaccines to 6 individuals each and measured the antibody presence in their blood after a chosen time period . The data is as follows:

| Vaccine | Antibodies (µg/ml) |
|---|---|
| A | 1232 |
| A | 751 |
| A | 339 |
| A | 848 |
| A | 447 |
| A | 542 |
| – | – |
| B | 302 |
| B | 57 |
| B | 521 |
| B | 278 |
| B | 176 |
| B | 201 |
| – | – |
| C | 839 |
| C | 342 |
| C | 473 |
| C | 1128 |

| Vaccine | Antibodies (µg/ml) |
|---|---|
| C | 242 |
| C | 475 |

*Ans: Here we see that the p-value is ~0.026 which is less than the cutoff 0.05, so we* ***reject the null hypothesis****: the medians are not the same across all three groups, at least one of them has a different median than the others. This means that the vaccines do not perform equally well because the resulting antibody production is not the same for each vaccine. We draw the same conclusion as we did above when we performed the calculation ourselves!*

*Again we emphasize that the Kruskal-Wallis test can only tell us that at least one of the vaccines performs differently than the others. It cannot tell us which vaccine(s) that is(are).*

13. The score of a sample of 20 students in their university examination are arranged according to the method used in their training : 1) Video Lectures 2) Books and Articles 3) Class Room Training. Evaluate the Effectiveness of these training methods at 0.10 level of significance.

| Video Lecture | Books and Articles | Class Room Training |
|---|---|---|
| 76 | 80 | 70 |
| 90 | 80 | 85 |
| 84 | 67 | 52 |
| 95 | 59 | 93 |

| Video Lecture | Books and Articles | Class Room Training |
|---|---|---|
| 57 | 91 | 86 |
| 72 | 94 | 79 |
| | 68 | 80 |

*Ans: Since, $H_{calc} < X^2$ . We* ***accept the Null Hypothesis****. We can say that there is no difference in the result obtained by using the three training methods.*

14. In a Study, 12 participants were divided into three groups of 4 each, they were subjected
to three different conditions, A (Low Noise), B(Avearge Noise), and C(Loud Noise). They were given a test and the errors committed by them on the test were noted and are given in the table below.

| Participant No. | Condition A (Low Noise) | Participant No. | Condition B (Average Noise) | Participant No. | Condition C (Loud Noise) |
|---|---|---|---|---|---|
| 1 | 3 | 5 | 2 | 9 | 10 |
| 2 | 5 | 6 | 7 | 10 | 8 |
| 3 | 6 | 7 | 9 | 11 | 7 |
| 4 | 3 | 8 | 8 | 12 | 11 |

*Ans: Since the critical value is more than the actual value we accept the null hypothesis that*
*all the three conditions A (Low Noise), B(Avearge Noise), and C(Loud Noise), do not differ from each other, therefore, in the said experiment there was no differences in the 1groups performance based on the noise level.*

15. A state court administrator asked the 24 court coordinators in the state's three largest counties to rate their relative need for training in case flow management on a Likert

scale (1 to 7).

1 = no training need

7 = critical training need

41

Training Need of Court Coordinators

| County A | County B | County C |
|---|---|---|
| 3 | 7 | 4 |
| 1 | 6 | 2 |
| 3 | 5 | 5 |
| 1 | 7 | 1 |
| 5 | 3 | 6 |
| 4 | 1 | 7 |
| 4 | 6 | |
| 2 | 4 | |
| | 4 | |
| | 5 | |

***Ans:*** *The critical chi-square table value of H for α = 0.05, and df = 2, is 5.991*

*Since 4.42 < 5.991, the null hypothesis is accepted. There is no difference in the training needs of the court coordinators in the three counties.*

16. Original data is displayed in the table below. Is there a difference between groups 1, 2 and 3 using alpha = 0.05?

| Gr-1 | Gr-2 | Gr-3 |
|---|---|---|
| 27 | 20 | 34 |
| 2 | 8 | 31 |
| 4 | 14 | 3 |
| 18 | 36 | 23 |
| 7 | 21 | 30 |
| 9 | 22 | 6 |

## Friedman Test

17. Department of Public health and safety monitors the measures taken to cleanup drinking water were effective. Trihalomethanes (THMs) at 12 counties drinking water compared before cleanup, 1 week later and 2 weeks after cleanup.

| County | Trihalomethanes (THMs) | | |
|---|---|---|---|
| | Before Cleanup | Week1 | Week2 |
| 1 | 21.1 | 19.2 | 18.4 |
| 2 | 24.1 | 22.3 | 21.2 |
| 3 | 14.1 | 12.9 | 12.9 |
| 4 | 18.1 | 17.8 | 17.3 |
| 5 | 15.4 | 15.1 | 14.9 |
| 6 | 16.2 | 15.1 | 15.1 |
| 7 | 7.4 | 7.2 | 6.8 |
| 8 | 7.5 | 6.7 | 6.1 |
| 9 | 14.2 | 13.6 | 13.1 |
| 10 | 21.3 | 20.9 | 20.4 |
| 11 | 9.5 | 9.8 | 9.2 |
| 12 | 11.9 | 10.5 | 10.1 |

*Ans: So, it is concluded that the cleanup system effected the THMs of drinking water.*

18. 7 random people were given 3 different drugs and for each person, the reaction time corresponding to the drugs were noted. Test the claim at the 5% significance level that all the 3 drugs have the same probability distribution.

| | Drug A | Drug B | Drug C |
|---|---|---|---|
| 1 | 1.24 | 1.50 | 1.62 |
| 2 | 1.71 | 1.85 | 2.05 |
| 3 | 1.37 | 2.12 | 1.68 |

| | Drug A | Drug B | Drug C |
|---|---|---|---|
| 4 | 2.53 | 1.87 | 2.62 |
| 5 | 1.23 | 1.34 | 1.51 |
| 6 | 1.94 | 2.33 | 2.86 |
| 7 | 1.72 | 1.43 | 2.86 |



*Ans: All the three drugs do not have the same probability distribution.*

19. Original data is displayed in the table below. Is there a difference between groups 1, 2 and 3 using alpha = 0.05?

**Ordinal data is displayed in the table below. Is there a difference between Weeks 1, 2, and 3 using alpha = 0.05?**

| Week 1 | Week 2 | Week 3 |
|---|---|---|
| 27 | 20 | 34 |
| 2 | 8 | 31 |
| 4 | 14 | 3 |
| 18 | 36 | 23 |
| 7 | 21 | 30 |
| 9 | 22 | 6 |

*Ans: There is no difference in the three gps*

# Module 6 – Practice Problems

1. Find the simple linear regression equation that fits the given data and coefficient of determination.

| Bill | Tip |
|---|---|
| 34 | 5 |
| 108 | 17 |
| 64 | 11 |
| 88 | 8 |
| 99 | 14 |
| 54 | 5 |

Answer: $y = 0.1462x - 0.8188$

coefficient of determination = $r^2 = 0.7493 = 74.93\%$

2. Find the simple linear regression equation that fits the given data and coefficient of determination.

| Hour | Temp |
|---|---|
| 2 | 21 |
| 4 | 27 |
| 6 | 29 |
| 8 | 86 |
| 10 | 86 |
| 12 | 92 |

Answer: $y = -3.533 + 8.1x$

coefficient of determination = $r^2 = 0.917 = 91.7\%$

3. Sales data of 10 months for a coffee house situated near a prime location of a city comprising the number of customers (in hundreds) and monthly sales (in Thousand Rupees) are given below:

| Sr | No. of customers (in hundreds) | Monthly sales (In Thousand Rs.) |
|---|---|---|
| 1 | 6 | 1 |
| 2 | 6.1 | 6 |
| 3 | 6.2 | 8 |
| 4 | 6.3 | 10 |
| 5 | 6.5 | 11 |
| 6 | 7.1 | 20 |
| 7 | 7.6 | 21 |
| 8 | 7.8 | 22 |
| 9 | 8 | 23 |
| 10 | 8.1 | 25 |

Find the simple linear regression equation and coefficient of determination that fits the given data.

Answer: y = -52.6 + 9.656x

4. A survey was conducted to relate the time required to deliver a proper presentation on a topic , to the performance of the student with the scores he/she receives. The following Table shows the matched data:

| Hours | Score |
|---|---|
| 0.5 | 57 |
| 0.75 | 64 |

| | |
|---|---|
| 1 | 59 |
| 1.25 | 68 |
| 1.5 | 74 |
| 1.75 | 76 |
| 2 | 79 |
| 2.25 | 83 |
| 2.5 | 85 |
| 2.75 | 86 |
| 3 | 88 |
| 3.25 | 89 |
| 3.5 | 90 |
| 3.75 | 94 |
| 4 | 96 |

Find the regression equation and coefficient of determination that will predict a student's score if we know how many hours the student studied.

Answer: y = 54.772 +10.857x

Coefficient of determination: 0.9460

5. Find the simple linear regression equation that fits the given data and coefficient of determination.

| X | Y |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 6 |
| 4 | 4 |
| 5 | 5 |

Answer: y = 2.2 + 0.6x

Coefficient of determination: 0.4091

6. Find the simple linear regression equation that fits the given data and coefficient of determination.

| X | Y |
|---|---|
| -2 | -1 |
| 1 | 1 |
| 3 | 2 |

Answer: y = 23/38x + 5/19

Coefficient of determination: 0.9944

7. Find the simple linear regression equation that fits the given data and coefficient of determination.

| X | Y |
|---|---|
| 0 | 2 |
| 1 | 3 |
| 2 | 5 |
| 3 | 4 |
| 4 | 6 |

Answer: y = 0.9x + 2.2

Coefficient of determination: 0.81

8. Find the simple linear regression equation that fits the given data and coefficient of determination.

| X | Y |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 3 | 5 |
| 4 | 7 |

Answer: y = 1.3x + 1.5

Coefficient of determination: 0.9657

9. Find the simple linear regression equation that fits the given data and coefficient of determination.

| X | Y |
|---|---|
| 2 | 69 |
| 9 | 98 |
| 5 | 82 |
| 5 | 77 |
| 3 | 71 |
| 7 | 84 |
| 1 | 55 |
| 8 | 94 |
| 6 | 84 |
| 2 | 64 |

Answer: y = 55.048 + 4.74x

Coefficient of determination: 0.9505

10. Effect of hours of mixing on temperature of wood pulp.

| X | Y |
|---|---|
| 2 | 21 |
| 4 | 27 |
| 6 | 29 |
| 8 | 64 |
| 10 | 86 |
| 12 | 92 |

Answer: y = 8.1x -3.533

Coefficient of determination : 0.917286

11. The following data

| $x$ | 1 | 20 | 30 | 40 |
|---|---|---|---|---|
| $y$ | 1 | 400 | 800 | 1300 |

is regressed with least squares regression to $y=a_0+a_1x$. The value of $a_1$ most nearly is

- 27.480
- 28.956
- 32.625
- 40.000

Answer: 32.625

12. An instructor gives the same $y$ vs $x$ data as given below to four students and asks them to regress the data with least squares regression to $y=a_0+a_1x$.

| | 1 | 10 | 20 | 30 | 40 |
|---|---|---|---|---|---|
| | 1 | 100 | 400 | 600 | 1200 |

Each student comes up with four different answers for the straight line regression model. Only one is correct. The correct model is

- $y=60x-1200$
- $y=30x-200$
- $y=-139.43+29.684x$
- $y=1+22.782x$

Answer: Y = -139.43+29.684x

The process of constructing a mathematical model or function that can be used to predict
or determine one variable by another variable is called

A. regression

B. correlation

C. residual

D. outlier plot

13. The process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable is called

A. regression B. correlation C. residual D. outlier plot

Ans: A

14. In the regression equation Y = 21 - 3X, the slope is

A. 21 B. -21 C. 3 D. -3

Ans: D

15. In the regression equation Y = 75.65 + 0.50X, the intercept is

A. 0.50 B. 75.65 C. 1.00 D. indeterminable

Ans: B

16. The difference between the actual Y value and the predicted Y value found using a regression equation is called the

A. slope B. residual C. outlier D. scatter plot

Ans: B

17. The total of the squared residuals is called the

A. coefficient of determination B. sum of squares of error C. standard error of the estimate D. r-squared

Ans: B

18. In regression analysis, $R^2$ is also called the

A. residual B. coefficient of correlation C. coefficient of determination D. standard error of the estimate

Ans: C

19. The coefficient of determination must be

A. between -1 and +1 B. between -1 and 0 C. between 0 and 1 D. equal to SSE/(n-2)

Ans: C

20. For a data set the regression equation is Y = 21 - 3X. The correlation coefficient for this data

A. must be 0 B. is negative C. must be 1 D. is positive

Ans: B

21. If X and Y in a regression model are totally unrelated,

A. the correlation coefficient would be -1 B. the coefficient of determination would be 0 C. the coefficient of determination would be 1 D. the SSE would be 0

Ans: B

**22-25 The following data is to be used to construct a regression model:**

| | | | | | | |
|---|---|---|---|---|---|---|
| X | 5 | 7 | 4 | 15 | 12 | 9 |
| Y | 8 | 9 | 12 | 26 | 16 | 13 |

22. The value of the intercept is

A. 1.36 B. 2.16 C. 0.68 D. 0.57

Ans: B

23. The value of the slope is for the data above is

A. 1.36 B. 2.16 C. 0.68 D. 0.57

Ans: A

24. The value of the coefficient of determination (R2) is

A. 0.78 B. 0.88 C. 0.36 D. 0.61

Ans: A

25. The value of the sum of squares of error (SSE) is

A. 11.85 B. 214.00 C. 47.39 D. 14.06

Ans: C

## MULTIPLE REGRESSION

26. In the context of Multiple linear regression explain what is Over fitting & multicollinearity?

Ans.

- Adding more independent variables to a multiple regression procedure does not mean the regression will be better of offer better predictions; in fact it can make things worse. This is called OVERFITTING.
- The addition of more independent variables creates more relationships among them. So not only are the independent variables potentially related to the dependent variable, they are also potentially related to each other. When this happens, it is called MULTICOLLINEARITY.

27. Predict equation for y.

| y | x1 | x2 |
|---|---|---|
| -3.7 | 3 | 8 |
| 3.5 | 4 | 5 |
| 2.5 | 5 | 7 |
| 11.5 | 6 | 3 |
| 5.7 | 2 | 1 |

Answer: $Y=2.8+2.28*X_1-1.67X_2$

28. Find out what is the relation between the distance covered by an UBER driver and the age of the driver and the number of years of experience of the driver.

| Distance | Age | Experience (in years) |
|---|---|---|
| 32513 | 18 | 5 |
| 27897 | 20 | 7 |
| 29929 | 22 | 8 |
| 20159 | 23 | 6 |
| 21554 | 23 | 7 |
| 28466 | 25 | 5 |
| 27842 | 2 | 8 |
| 22671 | 28 | 6 |
| 32214 | 29 | 5 |
| 34550 | 32 | 7 |
| 20920 | 37 | 9 |
| 33714 | 41 | 6 |
| 26998 | 46 | 7 |
| 34294 | 49 | 8 |
| 21912 | 53 | 6 |



Answer: The regression formula for the above example will be

Y=31216.5+(13.24*X1)-(585.46*X2)

In this particular example, we will see which variable is the dependent variable and which variable is the independent variable. The dependent variable in this regression equation is the distance covered by the UBER driver, and the independent variables are the age of the driver and the number of experiences he has in driving.

29. Find out what is the relation between the GPA of a class of students and the number of hours of study and the height of the students.

| GPA | Height | Study Hours |
|---|---|---|
| 2.9 | 66 | 7 |
| 3.16 | 57 | 7 |
| 3.62 | 64.5 | 6 |
| 2 | 62 | 7 |
| 3.45 | 69.5 | 8 |
| 2.8 | 65 | 9 |
| 3.63 | 63 | 6 |
| 2.81 | 68 | 5 |
| 3.33 | 59.5 | 4 |
| 2.75 | 64 | 10 |
| 3.86 | 69 | 7 |

Answer:

The regression equation for the above example will be

y=1.38+(0.038*X1)-(0.1*X2)

In this particular example, we will see which variable is the dependent variable and which variable is the independent variable. The dependent variable in this regression is the GPA, and the independent variables are study hours and height of the students.

30. Find out what is the relation between the salary of a group of employees in an organization and the number of years of experience and the age of the employees.

| Income | Age | Experience |
|---|---|---|
| 26315 | 18 | 5 |
| 39493 | 20 | 7 |
| 37209 | 22 | 8 |
| 24380 | 23 | 6 |
| 25751 | 23 | 7 |
| 44629 | 25 | 5 |
| 37616 | 2 | 8 |
| 33305 | 28 | 6 |
| 36848 | 29 | 5 |
| 42551 | 32 | 7 |
| 25700 | 37 | 9 |
| 37303 | 41 | 6 |
| 24659 | 46 | 7 |
| 32617 | 49 | 8 |
| 35771 | 53 | 6 |



Answer:

The regression equation for the above example will be

y=41350.4-(60.266*X1)-(891.1*X2)

In this particular example, we will see which variable is the dependent variable and which variable is the independent variable. The dependent variable in this regression equation is the salary, and the independent variables are the experience and age of the employees.