1. **Calculate covariance and correlation between below two columns A and B**

A	B
25	52
35	10
21	5
67	98
98	52
27	36
64	69

Mention all step by step formula calculations in the answer sheet.

Ans) A = 25,35,21,67,98,27,64

B = 52,10,5,98,52,36,69

Mean $\bar{A}$ = 48.1428

Mean $\bar{B}$ = 46

N = 7

Covariance(A,B) of the population = COV(A,B) = $\frac{\sum_{i=1}^{n}(Ai-\bar{A})(\mathrm{Bi}-\bar{\mathrm{B}})}{N}$

= (25-48.143)*(52-46)+(35-48.143)*(10-46)+(21-48.143)*(5-46)+(67-48.143)*(98-46)+(98-48.143)*(52-46)+(27-48.143)*(36-46)+(64-48.143)*(69-46))/7

= 471.8571

Covariance(A,B) of the sample = COV(A,B) = $\frac{\sum_{i=1}^{n}(Ai-\bar{A})(\mathrm{Bi}-\bar{\mathrm{B}})}{N-1}$

= (25-48.143)*(52-46)+(35-48.143)*(10-46)+(21-48.143)*(5-46)+(67-48.143)*(98-46)+(98-48.143)*(52-46)+(27-48.143)*(36-46)+(64-48.143)*(69-46))/(7-1)

= (25-48.143)*(52-46)+(35-48.143)*(10-46)+(21-48.143)*(5-46)+(67-48.143)*(98-46)+(98-48.143)*(52-46)+(27-48.143)*(36-46)+(64-48.143)*(69-46))/6

= 550.5

Variance(A) of the population = $\sigma^2 = \frac{\sum(A-\mu)^2}{N}$ =

$$\frac{((25-48.1428)+(35-48.1428)+(21-48.1428)+(67-48.1428)+(98-48.1428)+(27-48.1428)+(64-48.1428)}{}$$

= 712.1224

Standard Deviation(A) of the population = sd(A) = $\sqrt{712.1224}$ = 26.6856

Variance(B) of the population = $\sigma^2 = \frac{\sum(B-\mu)^2}{N}$ =

$$\frac{(52-46)+(10-46)+(5-46)+(98-46)+(52-46)+(36-46)+(69-46)}{7}$$

= 911.7143

Standard Deviation(B) of the population = sd(B) = $\sqrt{911.7143}$

= 30.1946

Correlation(A,B) of the population = $\frac{COV(A,B)}{sd(A)sd(B)}$

$$= \frac{471.8571}{26.6859*30.1946}$$

= 0.585604

Variance(A) of the sample = $\sigma^2 = \frac{\sum(A-\mu)^2}{N-1}$ =

$$\frac{((25-48.1428)+(35-48.1428)+(21-48.1428)+(67-48.1428)+(98-48.1428)+(27-48.1428)+(64-48.1428)}{7-1}$$

= 830.8095

Standard Deviation(A) of the sample = sd(A) = $\sqrt{830.8095}$ = 28.8237

Variance(B) of the sample = $\sigma^2 = \frac{\sum(B-\mu)^2}{N}$=

$$\frac{(52-46)+(10-46)+(5-46)+(98-46)+(52-46)+(36-46)+(69-46)}{7-1}$$

= 1063.667

Standard Deviation(B) of the sample = sd(B) = $\sqrt{1063.667}$

= 32.6139

Correlation(A,B) of the sample = $\frac{COV(A,B)}{sd(A)sd(B)}$

$$= \frac{550.5}{1063.667 * 32.6139}$$

= 0.0158

2. What are the different ways to deal with multi collinearity?

Ans) Ways of dealing with collinearity or multi collinearity

– Ignore it. If prediction of y values is the object of your study, then collinearity is not a problem. It means that if the dependent variable and the target variable is highly correlated then we need to consider both the variables and don't need to remove any variable. So that multi collinearity can be accepted. But there should not be a multi collinear effect between the dependent and the target variable

– Get rid of the redundant variables using a variable selection technique.

3. What should be the correlation threshold value based on which we determine the highly collinear variables?

Ans) Correlation is a scaled version of Covariance and values ranges from -1 to +1. If the correlation value is -1 then we can say that this is highly negative correlated to each other. If it is zero, then there is no correlation between the two variables. And if the value between two independent variables is +1 then they are highly correlated to each other.

So '1' should be the correlation threshold value based on which we determine the highly collinear variables.

4. **What are the two different types of variable we used in ANOVA?**

Ans) In ANOVA, we use one categorical variable and one numerical variable.

Categorical variables are names or labels such as color, gender, name, etc.

Numerical variables are only numeric values such as age, marks, etc.

5. **What are the null and alternate hypothesis in chi-square test?**

Ans)

- Null hypothesis: There are no relationships between the categorical variables. If the value of one variable is known, then it does not help us predict the value of another variable.
- Alternative hypothesis: There are relationships between the categorical variables. Knowing the value of one variable does help us predict the value of another variable.

For example, suppose we have some data on credit card spending by males and females. It has two categorical values: “Gender” and “Credit Card spending”. Further, the “Gender” has two categories, i.e. “Male” and “Female”. Similarly, “Credit Card spending” has three categories: “High”, “Medium”, and “Low”. Now let the null hypothesis be that these two categorical values (“Gender” and “Credit Card spending”) be independent and so the alternative hypothesis is dependent on each other. This means that irrespective of male and female, the credit card spending will be high, low and medium in case of null hypothesis. But in case of alternative hypothesis, if it is a male then the spending will be high and if it is a female then the spending will be low.