Research Methodology

Lesson – 7

Content

- Measure of Central Tendency
- Measures of Dispersion
- Measures of Dispersion - Range
- Measures of Dispersion – Quartile Deviation
- Measures of Dispersion – Mean Deviation
- Measures of Dispersion – Standard Deviation – Change formula
- Measures of Dispersion – Standard Deviation
- Normal Distribution
- Skewness and Kurtosis
- Pictorial Representation
- Standardizing Data
- Correlation

- Correlation - Graphical Representation
- Pearson Correlation
- Regression
- Regression – Assumptions
- Chi Square Analysis
- Chi Square – Application
- Chi Square Analysis - How Does It Work
- Chi Square – Steps
- Limitations of Chi Square Analysis
- ANOVA
- ANOVA - Calculation of the F-ratio
- Assumptions of ANOVA
- ANOVA – Types
- ANCOVA
- Links

Measure of Central Tendency

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

- The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

- Mean

  √ The Mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.

  √ The mean has one main disadvantage: it is particularly susceptible to the influence of outliers.

- Median
  - √ The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

- Mode
  - √ The mode is the most frequent score in our data set. Normally, the mode is used for categorical data where we wish to know which is the most common category.

Type of Variable	**Best measure of central tendency**
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Measures of Dispersion

- As the name suggests, the measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations. The main idea about the measure of dispersion is to get to know how the data are spread. It shows how much the data vary from their average value.

Measures of Dispersion - Range

- A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X_{max} and X_{min} are the two extreme observations then

- Range = $X_{max} - X_{min}$

- **Merits of Range**

  √ It is the simplest of the measure of dispersion
  √ Easy to calculate
  √ Easy to understand
  √ Independent of change of origin

- **Demerits of Range**

  √ It is based on two extreme observations. Hence, get affected by fluctuations
  √ A range is not a reliable measure of dispersion
  √ Dependent on change of scale

Measures of Dispersion – Quartile Deviation

- The quartiles divide a data set into quarters. The first quartile, (Q_1) is the middle number between the smallest number and the median of the data. The second quartile, (Q_2) is the median of the data set. The third quartile, (Q_3) is the middle number between the median and the largest number.
- Quartile deviation or semi-inter-quartile deviation is $Q = ½ \times (Q_3 – Q1)$
- **Merits of Quartile Deviation**
  - √ All the drawbacks of Range are overcome by quartile deviation
  - √ It uses half of the data
  - √ Independent of change of origin
  - √ The best measure of dispersion for open-end classification
- **Demerits of Quartile Deviation**
  - √ It ignores 50% of the data
  - √ Dependent on change of scale
  - √ Not a reliable measure of dispersion

Measures of Dispersion – Mean Deviation

- Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If $x_1, x_2, \dots, x_n$ are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is
- Mean deviation from average $A = 1/n\ [\sum_i |x_i - A|]$
- For a grouped frequency, it is calculated as:
- Mean deviation from average $A = 1/N\ [\sum_i f_i\ |x_i - A|]$, $N = \sum f_i$
- Here, x_i and f_i are respectively the mid value and the frequency of the i^{th} class interval.

- **Merits of Mean Deviation**
  - √ Based on all observations
  - √ It provides a minimum value when the deviations are taken from the median
  - √ Independent of change of origin

- **Demerits of Mean Deviation**
  - √ Not easily understandable
  - √ Its calculation is not easy and time-consuming
  - √ Dependent on the change of scale
  - √ Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment

Measures of Dispersion – Standard Deviation – Change formula

- A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation. The standard deviation is given as

$$\sigma = [(\Sigma_i (y_i - \bar{y})\, 2/\, n]^{½} = [(\Sigma_i y_i^{\,2} / n) - \bar{y}^{\,2}]^{½}$$

- For a grouped frequency distribution, it is

$$\sigma = [(\Sigma_i f_i (y_i - \bar{y}) / N]^{½} = [(\Sigma_i f_i y_i^{\,2} / n) - \bar{y}^{\,2}]^{½}$$

- The square of the standard deviation is the **variance**. It is also a measure of dispersion.

$$\sigma^{\,2} = [(\Sigma_i (y_i - \bar{y})2 / n] = [(\Sigma_i y_i^{\,2} / n) - \bar{y}^{\,2}]$$

$\bar{y}$ is the mid point of the class

- For a grouped frequency distribution, it is

$$\sigma^2 = [(\Sigma_i f_i (y_i - \bar{y}) / N]^{1/2} = [(\Sigma_i f_i x_i^2 / n) - \bar{y}^2].$$

- If instead of a mean, we choose any other arbitrary number, say A, the standard deviation becomes the root mean deviation.

Measures of Dispersion – Standard Deviation

- **Merits of Standard Deviation**
  - √ Squaring the deviations overcomes the drawback of ignoring signs in mean deviations
  - √ Suitable for further mathematical treatment
  - √ Least affected by the fluctuation of the observations
  - √ The standard deviation is zero if all the observations are constant
  - √ Independent of change of origin

- **Demerits of Standard Deviation**
  - √ Not easy to calculate
  - √ Difficult to understand for a layman
  - √ Dependent on the change of scale

Normal Distribution

- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve.
- A normal distribution is the proper term for a probability bell curve.
- In a normal distribution the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.
- Normal distributions are symmetrical , but not all symmetrical distributions are normal.
- In reality, most pricing distributions are not perfectly normal.
- For a normal distribution, 68% of the observations are within +/- one standard deviation of the mean, 95% are within +/- two standard deviations, and 99.7% are within +- three standard deviations.

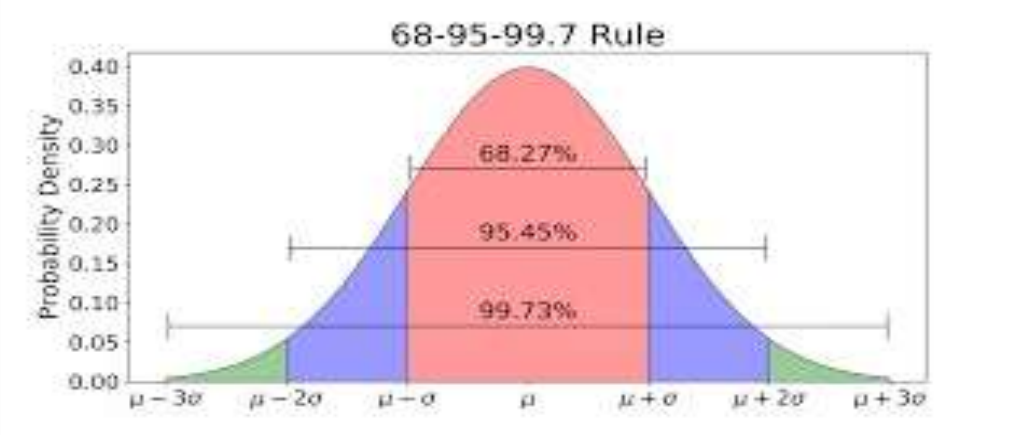
Skewness and Kurtosis

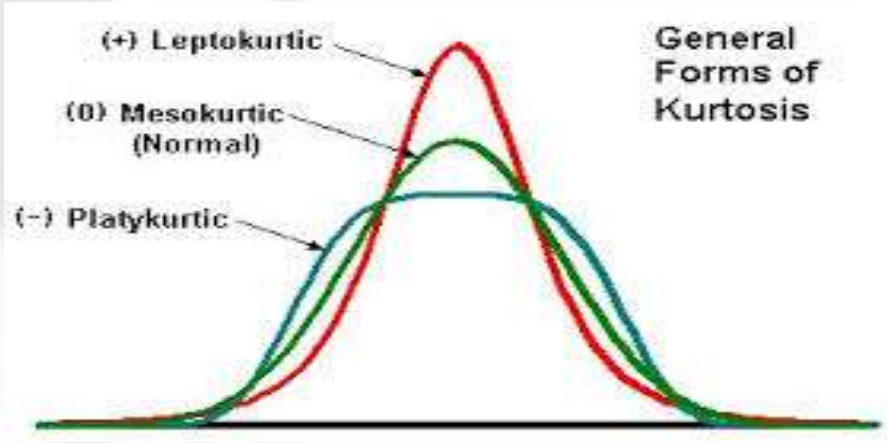
- The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution.

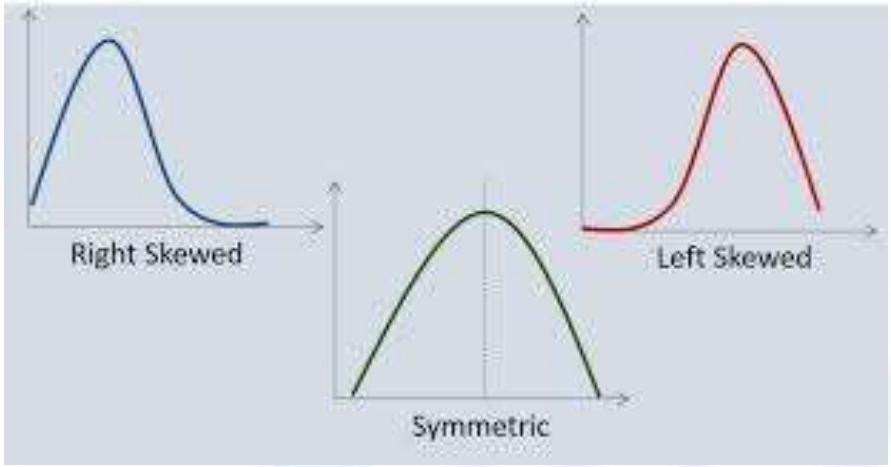
- The skewness measures the symmetry of a distribution. The normal distribution is symmetric and has a skewness of zero. If the distribution of a data set has a skewness less than zero, or negative skewness, then the left tail of the distribution is longer than the right tail; positive skewness implies that the right tail of the distribution is longer than the left.

- The kurtosis statistic measures the thickness of the tail ends of a distribution in relation to the tails of the normal distribution. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). Distributions with low kurtosis exhibit tail data that is generally less extreme than the tails of the normal distribution. The normal distribution has a kurtosis of three, which indicates the distribution has neither fat nor thin tails. Therefore, if an observed distribution has a kurtosis greater than three, the distribution is said to have heavy tails when compared to the normal distribution. If the distribution has a kurtosis of less than three, it is said to have thin tails when compared to the normal distribution.

Pictorial Representation







Standardizing Data

- The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score". Get used to those words!

- So to convert a value to a Standard Score ("z-score"):
  - √ first subtract the mean,
  - √ then divide by the Standard Deviation
  - √ And doing that is called "Standardizing":

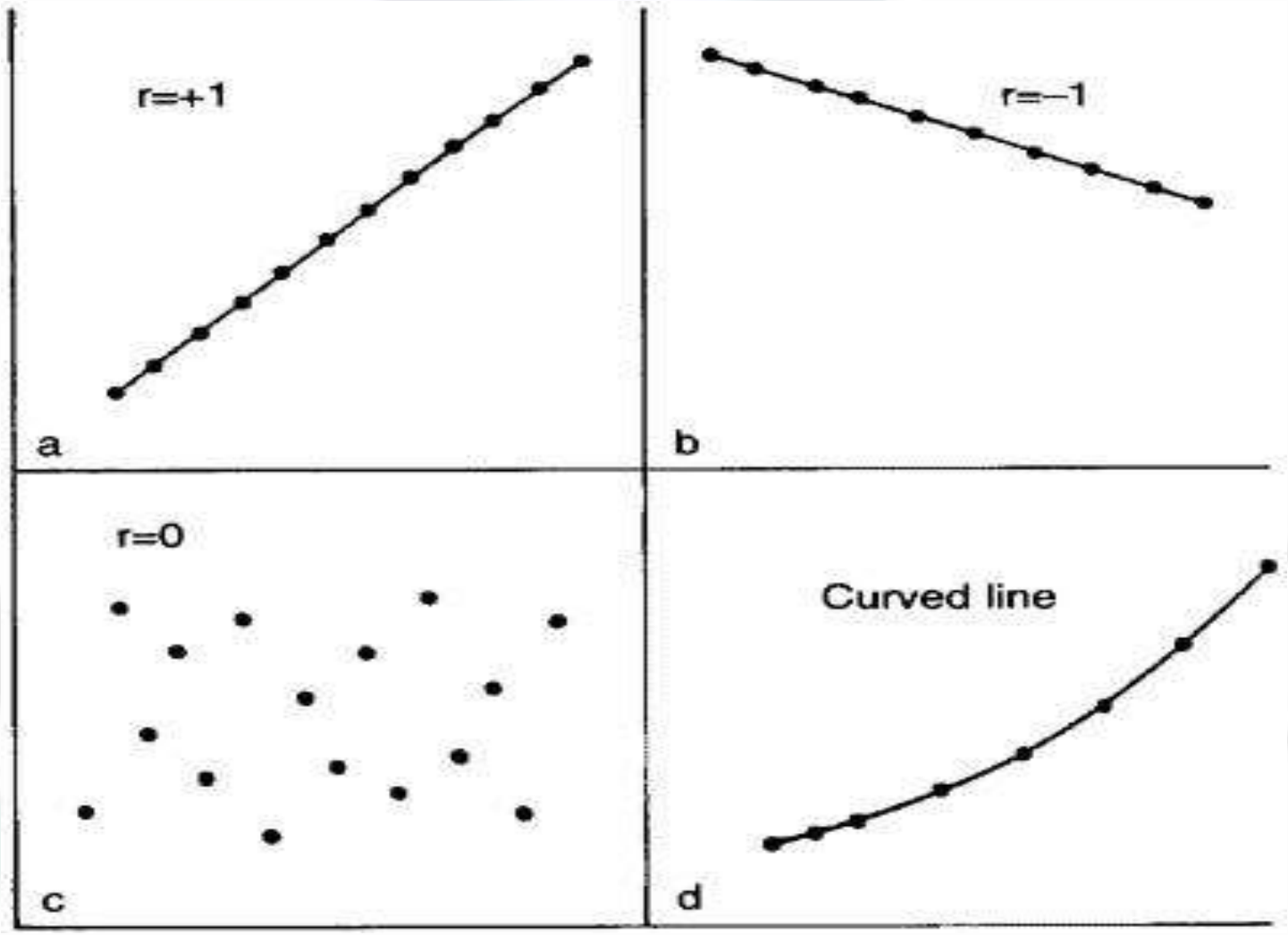
- The z-score formula that we have been using is:
  - ► $z = (x - \mu)/\sigma$
    - √ z is the "z-score" (Standard Score)
    - √ x is the value to be standardized
    - √ μ ('mu") is the mean
    - √ σ ("sigma") is the standard deviation

Correlation

- The word correlation is used in everyday life to denote some form of association. In statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other. The other technique that is often used in these circumstances is regression, which involves estimating the best straight line to summarise the association.

- The correlation coefficient is measured on a scale that varies from + 1 through 0 to - 1. Complete correlation between two variables is expressed by either + 1 or -1. When one variable increases as the other increases the correlation is positive; when one decreases as the other increases it is negative. Complete absence of correlation is represented by 0.

Correlation - Graphical Representation



Pearson Correlation

- Below is the formula for Pearson (most widely used) r correlation

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{[\Sigma(x - \bar{x})^2 (y - \bar{y})^2]}}$$

- For the Pearson r correlation, both variables should be normally distributed (normally distributed variables have a bell-shaped curve). Other assumptions include linearity and homoscedasticity. Linearity assumes a straight line relationship between each of the two variables and homoscedasticity assumes that data is equally distributed about the regression line.

Regression

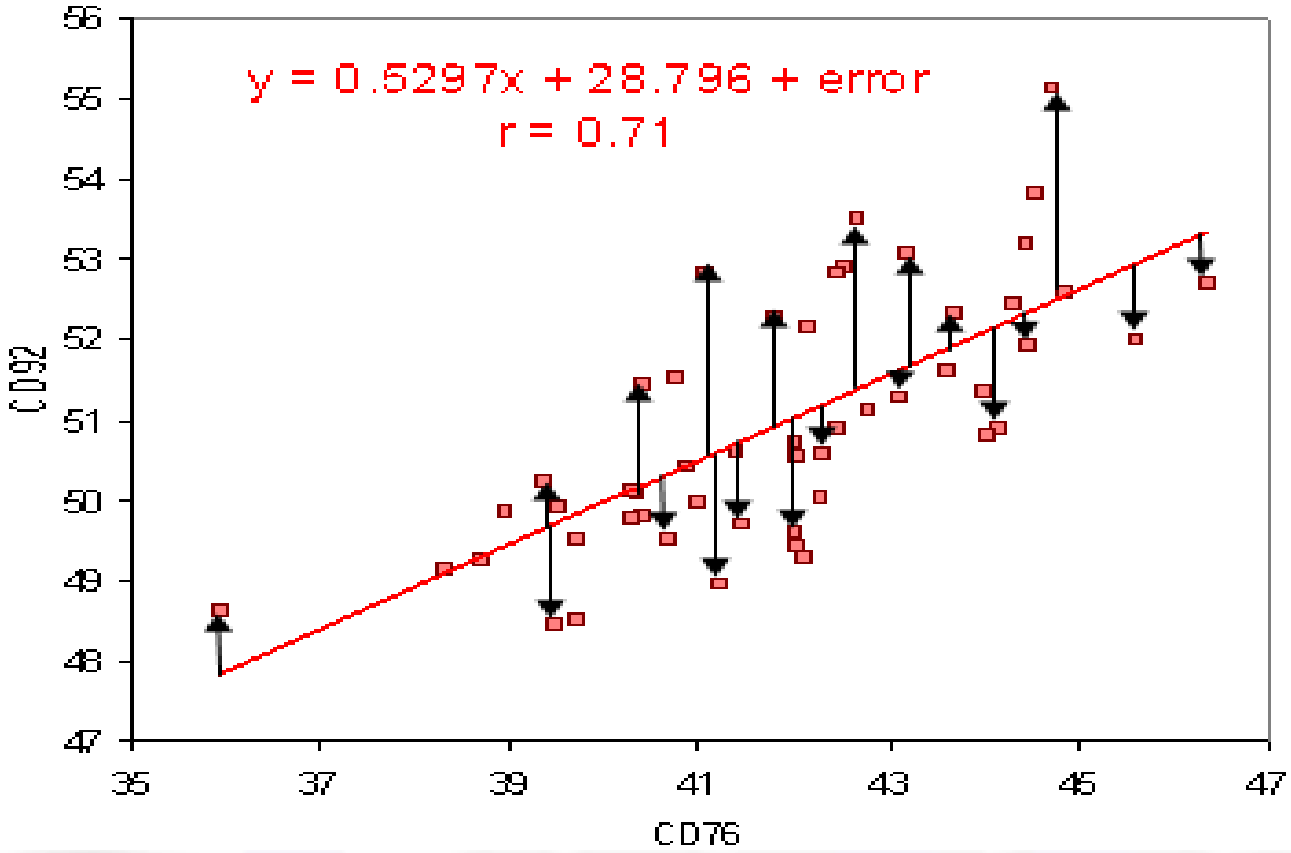
- Correlation describes the strength of an association between two variables, and is completely symmetrical, the correlation between x and y is the same as the correlation between y and x. However, if the two variables are related it means that when one changes by a certain amount the other changes on an average by a certain amount. If y represents the dependent variable and x the independent variable, this relationship is described as the regression of y on x.

- The relationship can be represented by a simple equation called the regression equation. In this context "regression" (the term is a historical anomaly) simply means that the average value of y is a "function" of x, that is, it changes with x. The simple linear model is expressed using the following equation:

$$Y=a+bX+\epsilon,$$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope
- **ϵ –** Residual (error)



r^2 is a measure of association; it represents the percent of the variance in the values of Y that can be explained by knowing the value of X. r^2 varies from a low of 0.0 (none of the variance is explained), to a high of +1.0 (all of the variance is explained).

The residual can be considered an estimate of the true error term

Regression – Assumptions

In theory, there are several important assumptions that must be satisfied if linear regression is to be used. These are:

1. Both the independent (X) and the dependent (Y) variables are measured at the interval or ratio level.
2. The relationship between the independent (X) and the dependent (Y) variables is linear.
3. Errors in prediction of the value of Y are distributed in a way that approaches the normal curve.
4. Errors in prediction of the value of Y are all independent of one another.
5. The distribution of the errors in prediction of the value of Y is constant regardless of the value of X.

Chi Square Analysis

- The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

- A chi-square ($\chi 2$) statistic is a test that measures expectations compare to actual observed data (or model results). The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample.

Chi Square – Application

- Chi-square test is a nonparametric test used for two specific purpose: (a) To test the hypothesis of no association between two or more groups, population or criteria (i.e. to check independence between two variables); (b) and to test how likely the observed distribution of data fits with the distribution that is expected (i.e., to test the goodness-of-fit). It is used to analyze categorical data (e.g. male or female patients, smokers and non-smokers, etc.), it is not meant to analyze parametric or continuous data (e.g., height measured in centimeters or weight measured in kg, etc.).

Chi Square Analysis - How Does It Work

The Chi-Square statistic is most commonly used to evaluate Tests of Independence when using a cross tabulation (also known as a bivariate table). Cross tabulation presents the distributions of two categorical variables simultaneously, with the intersections of the categories of the variables appearing in the cells of the table. The Test of Independence assesses whether an association exists between the two variables by comparing the observed pattern of responses in the cells to the pattern that would be expected if the variables were truly independent of each other. Calculating the Chi-Square statistic and comparing it against a critical value from the Chi-Square distribution allows the researcher to assess whether the observed cell counts are significantly different from the expected cell counts.

The calculation of the Chi-Square statistic is quite straight-forward and intuitive. The formula is shown on the right.

As depicted in the formula, the Chi-Square statistic is based on the difference between what is actually observed in the data and what would be expected if there was truly no relationship between the variables.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = Degrees of freedom

O = Observed value(s)

E = Expected value(s)

Chi Square – Steps

- **State the Hypotheses**
  - √ Suppose that Variable A has r levels, and Variable B has c levels. The null hypothesis states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.
    - Ho: Variable A and Variable B are independent.
    - Ha: Variable A and Variable B are not independent.

- Formulate an Analysis Plan
  - √ The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.
    - Significance level. Often, researchers choose significance levels equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
    - Test method. Use the chi-square test for independence to determine whether there is a significant relationship between two categorical variables.

- Analyse Sample Data

√ Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to: DF = (r - 1) * (c - 1)

where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

- **Expected frequencies.** The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute r * c expected frequencies, according to the following formula.

$$E_{r,c} = (n_r * n_c) / n$$

where $E_{r,c}$ is the expected frequency count for level *r* of Variable A and level *c* of Variable B, n_r is the total number of sample observations at level r of Variable A, n_c is the total number of sample observations at level *c* of Variable B, and n is the total sample size.

- **Test statistic.** The test statistic is a chi-square random variable (X^2) defined by the following equation.

$$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

where $O_{r,c}$ is the observed frequency count at level *r* of Variable A and level *c* of Variable B, and $E_{r,c}$ is the expected frequency count at level *r* of Variable A and level *c* of Variable B.

- **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedom computed above.

- **Interpret Results**

√ If the sample findings are unlikely, given the null hypothesis, the researcher rejects the null hypothesis. Typically, this involves comparing the P-value to the significance level, and rejecting the null hypothesis when the P-value is less than the significance level.

Limitations of Chi Square Analysis

- First, chi-square is highly sensitive to sample size. As sample size increases, absolute differences become a smaller and smaller proportion of the expected value. What this means is that a reasonably strong association may not come up as significant if the sample size is small, and conversely, in large samples, we may find statistical significance when the findings are small and uninteresting., i.e., the findings are not substantively significant, although they are statistically significant.

- Chi-square is also sensitive to small frequencies in the cells of tables. Generally when the expected frequency in a cell of a table is less than 5, chi-square can lead to erroneous conclusions. The rule of thumb here is that if either (i) an expected value in a cell is less than 5 or (ii) more than 20% of the expected values in cells are less than 5, then chi-square should not and usually is not computed.

ANOVA

- ANOVA is used to compare differences of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found (hence its name). Specifically, ANOVA compares the amount of variation between groups with the amount of variation within groups. It can be used for both observational and experimental studies.

- Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples.

- **Hypothesis testing**

  √ Like other classical statistical tests, we use ANOVA to calculate a test statistic (the F-ratio) with which we can obtain the probability (the P-value) of obtaining the data assuming the null hypothesis. A significant P-value (usually taken as $P<0.05$) suggests that at least one group mean is significantly different from the others.

  - *Null hypothesis:* all population means are equal
  - *Alternative hypothesis:* at least one population mean is different from the rest.

ANOVA - Calculation of the F-ratio

- ANOVA separates the variation in the dataset into 2 parts: between-group and within-group. These variations are called the sums of squares, which can be seen in the equations below.
- Step 1) Variation between groups
- The between-group variation (or between-group sums of squares, SS) is calculated by comparing the mean of each group with the overall mean of the data.
- Specifically, this is:
- i.e., by adding up the square of the differences between each group mean and the overall population mean , multiplied by sample size , assuming we are comparing three groups (i = 1, 2 or 3).
- We then divide the BSS by the number of degrees of freedom [this is like sample size, except it is n-1, because the deviations must sum to zero, and once you know n-1, the last one is also known] to get our estimate of the mean variation between groups.

- *Step 2) Variation within groups*
- The within-group variation (or the within-group sums of squares) is the variation of each observation from its group mean.

$$SS_R = s^2_{group1}\,(n_{group1} - 1) + s^2_{group2}\,(n_{group2} - 1) + s^2_{group3}\,(n_{group3} - 1)$$

- i.e., by adding up the variance of each group times by the degrees of freedom of each group. Note, you might also come across the total SS (sum of $(n_1 - x)^2$. Within SS is then Total SS minus Between SS.
- As before, we then divide by the total degrees of freedom to get the mean variation within groups.
- Step 3) The F ratio is then calculated as: $\frac{\text{Mean Between-group SS}}{\text{Mean Within-group SS}}$
- If the average difference between groups is similar to that within groups, the F ratio is about 1. As the average difference between groups becomes greater than that within groups, the F ratio becomes larger than 1.
- To obtain a P-value, it can be tested against the F-distribution of a random variable with the degrees of freedom associated with the numerator and denominator of the ratio. The P-value is the probably of getting that F ratio or a greater one. Larger F-ratios gives smaller P-values.

Assumptions of ANOVA

- The response is normally distributed
- Variance is similar within different groups
- The data points are independent

ANOVA – Types

- There are two main types of ANOVA: (1) "one-way" ANOVA compares levels (i.e. groups) of a single factor based on single continuous response variable (e.g. comparing test score by 'level of education') and (2) a "two-way" ANOVA compares levels of two or more factors for mean differences on a single continuous response variable(e.g. comparing test score by both 'level of education' and 'zodiac sign'). In practice, you will see one-way ANOVAs more often and when the term ANOVA is generically used, it often refers to a one-way ANOVA.

ANCOVA

- The obvious difference between ANOVA and ANCOVA is the the letter "C", which stands for 'covariance'. Like ANOVA, "Analysis of Covariance" (ANCOVA) has a single continuous response variable. Unlike ANOVA, ANCOVA compares a response (dependent) variable by both a factor and a continuous independent variable (e.g. comparing test score by both 'level of education' and 'number of hours spent studying'). The term for the continuous independent variable (IV) used in ANCOVA is "covariate".

Links

- https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests#:~:text=Parametric%20tests%20are%20those%20that,used%20for%20non%2DNormal%20variables.
- (Above Link Describes Various Parametric (and non) Tests
- https://stattrek.com/chi-square-test/independence.aspx
- https://www.investopedia.com/terms/c/chi-square-statistic.asp

End of Lesson 7