ADM COLLEGE FOR WOMEN (A), NAGAPATTINAM
PG & RESEARCH DEPARTMENT OF ECONOMICS
SUBJECT : STATISTICS FOR ECONOMICS - II
CLASS : I BA ECONOMICS
HANDLED BY Dr.R.L.POONGUZHALI

UNIT - I

INTRODUCTION

Index number is a technique of measuring changes in a variable or a group of variables with respect to time, location or other characteristics. It is one of the most widely used statistical methods. Index number is a specialized average designed to measure the change in a group of related variables over a period of time. For example, the price of cotton in 2010 is studied with reference to its price in 2000. It is used to feel the pulse of the economy and it reveals the inflationary or deflationary tendencies. In reality, it is viewed as barometers of economic activity because if one wants to have an idea as to what is happening in an economy, he should check the important indicators like the index number of agricultural production, index number of industrial production, and the index number business activity etc.,

INDEX NUMBER MEANING

An index number is a method of evaluating variations in a variable or group of variables in regards to geographical location, time, and other features. The base value of the index number is usually 100, which indicates price, date, level of production, and more

DEFINITION

An Index Number is defined as a relative measure to compare and describe the average change in price, quantity value of an item or a group of related items with respect to time, geographic location or other characteristics accordingly.

In the words of **Maslow** "An index number is a numerical value characterizing the change in complex economic phenomenon over a period of time or space"

Spiegal defines, "An index number is a statistical measure designed to show changes in a variable on a group of related variables with respect to time, geographical location or other characteristics".

According to **Croxton and Cowden** "Index numbers are devices for measuring differences in the magnitude of a group of related variables".

Bowley describes "Index Numbers as a series which reflects in its trend and fluctuations the movements of some quantity".

USES OF INDEX NUMBERS

The various uses of index numbers are:

➢ **Economic Parameters**

- The Index Numbers are one of the most useful devices to know the pulse of the economy.
- It is used as an indicator of inflanationary or deflanationary tendencies.

➢ **Measures Trends**

- Index numbers are widely used for measuring relative changes over successive periods of time.
- This enable us to determine the general tendency.
- For example, changes in levels of prices, population, production etc. over a period of time are analysed.

➢ **Useful for comparsion**

- The index numbers are given in percentages.
- So it is useful for comparison and easy to understand the changes between two points of time.

➢ **Help in framing suitable policies**

- Index numbers are more useful to frame economic and business policies.
- For example, consumer price index numbers are useful in fixing dearness allowance to the employees.

➢ **Useful in deating**

- Price index numbers are used for connecting the original data for changes in prices.
- The price index are used to determine the purchasing power of monetary unit.

➢ **Compares standard of living**

- Cost of living index of different periods and of different places will help us to compare the standard of living of the people.
- The enables the government to take suitable welfare measures.

➢ **Special type of average**

- All the basic ideas of averages are employed for the construction of index numbers.
- In averages, the data are homogeneous (in the same units) but in index number, we average the variables which have different units of measurements.
- Hence, it is a special type of average.

TYPES OF INDEX NUMBERS

(i) Price Index Numbers

Price index is a ‘Special type’ of average which studies net relative change in the prices of commodities, expressed in different units. Here comparison is made in respect of prices. Price index numbers are wholesale price index numbers and retail price index numbers.

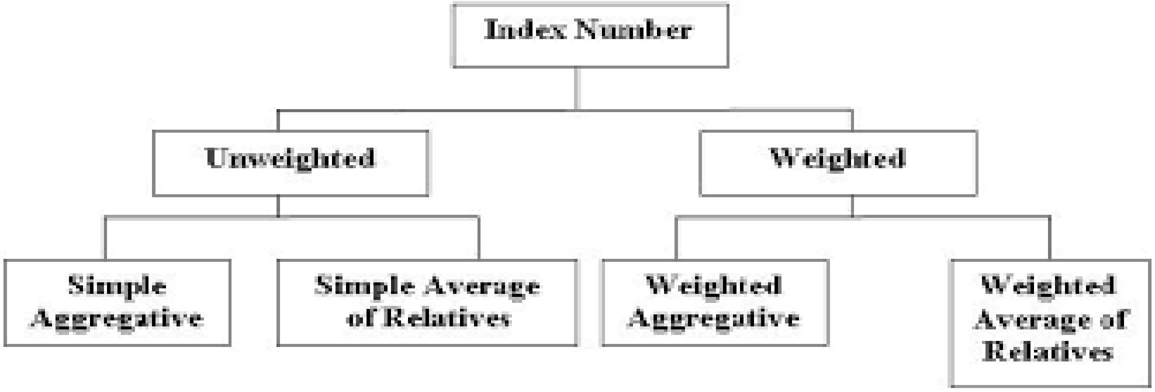
(ii) Quantity Index Numbers

This number measures changes in volume of goods produced, purchased or consumed. Here, the comparison is made in respect of quantity or volume. For example, the volume of agricultural goods produced, consumed, import, export etc.

(iii) Value Index

Value index numbers study the changes in the total value of a certain period with the total value of the base period. For example, the indices of stock-in-made, purchase, sales profit etc., are analysed here.

METHODS OF CONSTRUCTING INDEX NUMBERS



UNWEIGHTED INDEX NUMBERS

Unweighted Index Numbers An unweighted price Index Number measures the percentage change in price of a single item or a group of items between two periods of time. In unweighted index numbers, all the values taken for study are of equal importance. There are two methods in this category. (i) Simple aggregative method: Under this method the prices of different items of current year are added and the total is divided by the sum of prices of the base year items and multiplied by 100.

$$P_{01} = \frac{\sum p1}{\sum p0} \times 100$$

p1 = Current year prices for various commodities

p0 = Base year prices for various commodities

P01 = Price Index number Limitations of the simple aggregative method

(i) Relative importance of the commodities is not taken into account.

(ii) Highly priced items influence the index number

IMPORTANT QUESTIONS

1. What is a time series?
2. What are the components of a time series?
3. Name different methods of estimating the trend?
4. Write short notes on irregular variation.

5. Mention the methods used to find seasonal indices?
6. What are the demerits of moving averages?
7. What are the merits of method of least squares?
8. Write the normal equations used in method of least squares?
9. Define forecasting.
10. What are the three types of forecasting?
11. What is a short-term forecast?
12. Write the uses of time series.
13. Explain semi-averages method
14. Write the merits of moving averages.
15. What is cyclical variation?
16. What is seasonal variation?
17. What are medium-term and long-term forecasts?
18. Describe the method of finding seasonal indices.
19. With what characteristic component of a time series should each of the following be associated.
    (i) An upturn in business activity
    (ii) Fire loss in a factory
    (iii) General increase in the sale of Television sets.

WEIGHTED INDEX NUMBERS

In computing weighted Index Numbers, the weights are assigned to the items to bring out their economic importance. Generally quanties consumed or value are used as weights. Weighted index numbers are also of two types

(i) Weighted aggregative

(ii) Weighted average of price relatives

WEIGHTED AGGREGATE INDEX NUMBERS

In this method price of each commodity is weighted by the quantity sale either in the base year or in the current year. There are various methods of assigning weights and thus there are many methods of constructing index numbers. Some of the important formulae used under this methods are

a) Laspeyre’s Index (P_{01}^{L})

b) Paasche’s Index (P_{01}^{P})

c) Dorbish and Bowley’s Index (P_{01}^{DB})

d) Fisher’s Ideal Index (P_{01}^{F})

e) Marshall-Edgeworth Index (P_{01}^{Em})

f) Kelly’s Index (P_{01}^{K})

1] SIMPLE AVERAGE OR PRICE RELATIVES METHOD

In this method, we find out the price relative of individual items and average out the individual values. Price relative refers to the percentage ratio of the value of a variable in the current year to its value in the year chosen as the base.

$$\text{Price relative (R)} = (P1 \div P2) \times 100$$

Here, P1= Current year value of item with respect to the variable and P2= Base year value of the item with respect to the variable. Effectively, the formula for index number according to this method is:

$$P = \sum[(P1 \div P2) \times 100] \div N$$

Here, N= Number of goods and P= Index number.

2] SIMPLE AGGREGATIVE METHOD

It calculates the percentage ratio between the aggregate of the prices of all commodities in the current year and aggregate prices of all commodities in the base year.

$$P= (\sum P1 \div \sum P2)\times 100$$

Here, $\sum P1$= Summation of the prices of all commodities in current year and $\sum P2$= Summation of prices of all commodities in base year.

What is a fixed and chain based index number?

A chain index is an index number in which the value of any given period is related to the value of its immediately preceding period (resulting in an index for the given period expressed against the preceding period = 100); this is distinct from the fixed-base index, where the value of every period in a time series is directly related to the same value of one fixed base period.

TESTS OF ADEQUACY OF INDEX NUMBERS:

We have seen that there are a number of formulae for constructing index numbers. Thus the problem would be to select a proper formula. The following four tests are suggested to compare the adequacy of a formula.

(1) Unit Test
(2) Time Reversal Test,
(3) Factor Reversal Test,
(4) Circular Test.

1. **Unit Test:** It is natural natural to expect that the index number should be free from units of measurement of quantities and the units of prices. All the above formulae, except the aggregative index formula satisfy this test. In this respect all of them are equally good.

2. **Time Reversal Test:** It states that if we calculate first P01 and then P10 by interchanging base year and the current year the product of the principal factors in the two index numbers should be equal to unity.

3. **Factor Reversal Test :** It states that the product of the factors of the index number of price and the index number of quantity should be equal to value ratio.

4. **Circular Test:** Another test of adequacy of index number is called circular test. It is a sort of extension of time reversal test.

WHOLESALE PRICE INDEX MEANING

Wholesale Price Index is a measure of the average change in the price of goods at a wholesale level or in the wholesale market.

DEFINITION

The Wholesale Price Index is the price of a representative basket of wholesale goods. Some countries use WPI changes as a central measure of inflation. But now India has adopted new CPI to measure inflation.

IMPORTANCE OF WHOLESALE PRICE INDEX

- In a dynamic world, prices do not remain constant.
- The inflation rate calculated on the basis of the movement of the Wholesale Price Index (WPI) is an important measure to monitor the dynamic movement of prices.
- As WPI captures price movements in a most comprehensive way, it is widely used by Government, banks, industry and business circles.
- Significant monetary and fiscal policy changes are often linked to WPI movements.
- Similarly, the movement of WPI serves as an influential determinant, in the formulation of trade, fiscal and other economic policies by the Government of India.
- The WPI indices are also used for the purpose of escalation clauses in the supply of raw materials, machinery and construction work.
- WPI is used as a deflator of various nominal macroeconomic variables, including Gross Domestic Product (GDP).

CONSUMER PRICE INDEX MEANING

Consumer Price Index is another price index that calculates price changes of goods and services that a consumer has to pay for consuming a basket of goods.

DEFINITION

Consumer Price Index or CPI is another price index that focuses on the sum of money that a consumer has to shell out in order to purchase a basket of goods and services over a period of time. CPI measures the price that consumers pay to retailers.

IMPORTANCE OF CONSUMER PRICE INDEX

- CPI is a widely used measure for determining inflation in an economy.

- Rising inflation results in the diminishing standard of living for the residents of a nation.
- Over a period of time, it will result in an increase in the cost of living.
- A high inflation rate will result in increase in prices of goods and as a result there will be less manufacturing, which will result in loss of jobs.

USES OF THE CONSUMER PRICE INDEX

- It serves as an indicator of inflation in an economy.
- Can be used to change the components of national income

LIMITATIONS OF CONSUMER PRICE INDEX

- CPI cannot calculate the variations in two different areas.
- It is a mechanism that detects conditional cost of living and not includes all aspects that impact the living standards.

UNIT - II

TIME SERIES ANALYSIS DEFINITION

Time series refers to an arrangement and presentation of statistical data in chronological order. The statistical data is collected over a period of time.

COMPONENTS OF TIME SERIES ANALYSIS

- Secular Trend or Simple trend or Long term movement
- Seasonal variations
- Cyclical variations
- Random or irregular variations

USES OD TIME SERIES ANALYSIS

- The most important use of studying time series is that it helps us to predict the future behaviour of the variable based on past experience
- It is helpful for business planning as it helps in comparing the actual current performance with the expected one
- From time series, we get to study the past behaviour of the phenomenon or the variable under consideration
- We can compare the changes in the values of different variables at different times or places, etc.

ADVANTAGES OF TIME SERIES ANALYSIS

- **Reliability:** Time series analysis uses historical data to represent conditions along with a progressive linear chart. The information or data used is collected over a period of time say, weekly, monthly, quarterly or annually. This makes the data and forecasts reliable.

- **Seasonal Patterns:** As the data related to a series of periods, it helps us to understand and predict the seasonal pattern. For example, the time series may reveal that the demand for ethnic clothes not only increases during Diwali but also during the wedding season.
- **Estimation of trends:** The time series analysis helps in the identification of trends. The data tendencies are useful to managers as they show an increase or decrease in sales, production, share prices, etc.
- **Growth**: Time series analysis helps in the measurement of financial growth. It also helps in measuring the internal growth of an organization that leads to economic growth.

DISADVANTAGES OF TIME SERIES ANALYSIS

- Time series analysis is not perfect.
- It can suffer from generalization from a single study where more data points and models were warranted.
- Human error could misidentify the correct data model, which can have a snowballing effect on the output.
- It could also be difficult to obtain the appropriate data points.
- A major point of difference between time-series analysis and most other statistical problems is that in a time series, observations are not always independent.

MEASUREMENTS OF TRENDS

Following are the methods by which we can measure the trend.
(i) Freehand or Graphic Method.
(ii) Method of Semi-Averages.
(iii) Method of Moving Averages.
(iv) Method of Least Squares.

(i) Freehand or Graphic Method.

It is the simplest and most flexible method for estimating a trend. We will see the working procedure of this method.

Procedure:

(a) Plot the time series data on a graph.
(b) Draw a freehand smooth curve joining the plotted points.
(c) Examine the direction of the trend based on the plotted points.
(d) Draw a straight line which will pass through the maximum number of plotted points.

(ii) Method of Semi-Averages

In this method, the semi-averages are calculated to find out the trend values. Now, we will see the working procedure of this method.

Procedure:

(i) The data is divided into two equal parts. In case of odd number of data, two equal parts can be made simply by omitting the middle year.

(ii) The average of each part is calculated, thus we get two points.

(iii) Each point is plotted at the mid-point (year) of each half.

(iv) Join the two points by a straight line.

(v) The straight line can be extended on either side.

(vi) This line is the trend line by the methods of semi-averages.

(iii) Method of Moving Averages

Moving Averages Method gives a trend with a fair degree of accuracy. In this method, we take arithmetic mean of the values for a certain time span. The time span can be three-years, four -years, five- years and so on depending on the data set and our interest. We will see the working procedure of this method.

Procedure:

(i) Decide the period of moving averages (three- years, four -years).

(ii) In case of odd years, averages can be obtained by calculating,

$$\frac{a+b+c}{3},\frac{b+c+d}{3},\frac{c+d+e}{3},\frac{d+e+f}{3},\dots\dots$$

(iii) If the moving average is an odd number, there is no problem of centering it, the average value will be centered besides the second year for every three years.

(iv) In case of even years, averages can be obtained by calculating,

$$\frac{a+b+c+d}{4},\frac{b+c+d+e}{4},\frac{c+d+e+f}{4},\frac{d+e+f+g}{4},\dots$$

(v) If the moving average is an even number, the average of first four values will be placed between 2 nd and 3rd year, similarly the average of the second four values will be placed between 3rd and 4th year. These two averages will be again averaged and placed in the 3rd year. This continues for rest of the values in the problem. This process is called as centering of the averages.

(iv) Method of Least Squares

The line of best fit is a line from which the sum of the deviations of various points is zero. This is the best method for obtaining the trend values. It gives a convenient basis for calculating the line of best fit for the time series. It is a mathematical method for measuring trend. Further the sum of the squares of these deviations would be least when compared with other fitting methods. So, this method is known as the Method of Least Squares and satisfies the following conditions:

(i) The sum of the deviations of the actual values of Y and $\hat{Y}$ (estimated value of Y) is Zero. that is $\Sigma(Y–\hat{Y}) = 0$.

(ii) The sum of squares of the deviations of the actual values of Y and $\hat{Y}$ (estimated value of Y) is least. that is $\Sigma(Y–\hat{Y})2$ is least ;

Procedure:

(i) The straight line trend is represented by the equation Y = a + bX …(1)
where Y is the actual value, X is time, a, b are constants
(ii) The constants 'a' and 'b' are estimated by solving the following two normal
Equations ΣY = n a + b ΣX ...(2)
ΣXY = a ΣX + b ΣX2 ...(3)
Where 'n' = number of years given in the data.
(iii) By taking the mid-point of the time as the origin, we get $\Sigma X = 0$
(iv) When $\Sigma X = 0$, the two normal equations reduces to

$$\Sigma Y = n\,a + b\,(0) \quad ; a = \frac{\sum Y}{n} = \bar{Y}$$

$$\Sigma XY = a(0) + b\,\Sigma X^2 \;; b = \frac{\sum XY}{\sum X^2}$$

The constant 'a' gives the mean of Y and 'b' gives the rate of change (slope).
(v) By substituting the values of 'a' and 'b' in the trend equation (1), we get the Line of Best Fit.

GRAPHICAL METHOD MEANING

Graphical methods are useful aids to portray the results of formal statistical tests of trends. In general, the formal test procedures can be viewed as methods that assign a probability level to the validity of the trends observed in graphs. Hence, we encourage the use of graphics to display time series

TYPES OF GRAPHICAL METHOD

- Line Graphs
- Bar Charts
- Pie Carts
- Scatter Plots
- Heat Maps
- Histograms
- Network Diagrams
- Box Plots

APPLICATIONS OF GRAPHICAL METHOD

- Business
- Social sciences

- Education
- Healthcare
- Sports

EXAMPLES OF GRAPHICAL METHOD

- Stock Market
- Social Media Analytics
- Traffic Analysis
- Medical Diagnostics
- Cybersecurity

HOW TO USE GRAPHICAL METHODS ?

- Identify the research question
- Collect and organize the data
- Select the appropriate graph
- Create the graph
- Analyze the graph
- Draw conclusions
- Iterate and refine

WHEN TO USE GRAPHICAL METHODS ?

- To identify patterns and trends
- To compare data
- To summarize data
- To communicate data
- To identify outliers

CHARACTERISTICS OF GRAPHICAL METHOD

- Visual Representation
- Simplicity
- Comparability
- Flexibility
- Accuracy
- Clarity

ADVANTAGES OF GRAPHICAL METHODS

- Clear visualization
- Effective comparison
- Improved decision-making
- Increased engagement

- Better understanding

LIMITATIONS OF GRAPHICAL METHODS

- Misleading representation
- Limited scope
- Time-consuming
- Technical skills
- Interpretation
- Accessibility

METHODS OF SEMI AVERAGE MEANING

By semi-averages is meant the averages of the two halves of a series. In this method, thus, the given series is divided into two equal parts (halves) and the arithmetic mean of the values of each part (half) is calculated. The computed means are termed as semi-averages.

ADVANTAGES

- This method is simple to understand as compare to other methods for measuring the secular trends.
- Everyone who applies this method will get the same result.

DISADVANTAGES

- The method assumes a straight line relationship between the plotted points without considering the fact whether that relationship exists or not.
- If we add more data to the original data then we have to do the complete process again for the new data to get the trend values and the trend line also changes.

MOVING AVERAGES MEANING

A time series is broadly classified into three categories of long-term fluctuations, short-term or periodic fluctuations, and random variations. A long-term variation or a trend shows the general tendency of the data to increase or decrease during a long period of time. The variation may be gradual but it is inevitably present.

ANALYSIS OF MOVING AVERAGES

- To identify the components and the net effect of whose interaction is shown by the movement of a time series, and
- To isolate, study, analyze and measure each component independently by making others constant.

MERITS OF MOVING AVERAGE METHOD

- Moving averages help in identifying the trends. This allows the traders to avail of and understand the trends established in the market.
- It also acts as a support system as it helps in determining potential price support.
- It provides the support to measure the momentum as well. It helps to determine the direction and strength of the asset's momentum.

DRAWBACKS OF MOVING AVERAGE

- The main problem is to determine the extent of the moving average which completely eliminates the oscillatory fluctuations.
- This method assumes that the trend is linear but it is not always the case.
- It does not provide the trend values for all the terms.
- This method cannot be used for forecasting future trend which is the main objective of the time series analysis.

LEAST SQUARE METHOD MEANING

The least square method is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve.

DEFINITION

The least-squares method is a crucial statistical method that is practised to find a regression line or a best-fit line for the given pattern. This method is described by an equation with specific parameters. The method of least squares is generously used in evaluation and regression.

LIMITATIONS OF LEAST SQUARE METHOD

- In the process of regression analysis, which utilizes the least-square method for curve fitting, it is inevitably assumed that the errors in the independent variable are negligible or zero.
- In such cases, when independent variable errors are non-negligible, the models are subjected to measurement errors.
- Therefore, here, the least square method may even lead to hypothesis testing, where parameter estimates and confidence intervals are taken into consideration due to the presence of errors occurring in the independent variables.

PROS

- Easy to apply and understand
- Highlights relationship between two variables
- Can be used to make predictions about future performance

CONS

- Only highlights relationship between two variables
- Doesn't account for outliers
- May be skewed if data isn't evenly distributed

TIME SERIES MEANING

Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time. Using data visualizations, business users can see seasonal trends and dig deeper into why these trends occur. With modern analytics platforms, these visualizations can go far beyond line graphs.

USES OF TIME SERIES ANALYSIS

- Time series helps in understanding the past behaviour of the variable under study.
- Time series helps in determining the type and nature of the changes in the given data.
- Time series analysis facilities comparison between two or more variables over a period of time.
- Time series are very much useful in predicting probable future activities based on past performance.
- Time series help in controlling current accomplishments by comparing actual performance with forecasted performance.

UNIT - III

PROBABILITY MEANING

Probability denotes the possibility of the outcome of any random event. The meaning of this term is to check the extent to which any event is likely to happen.

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

P(E) = Number of Favourable Outcomes/Number of total outcomes

n(E) = Number of event favourable to event E

n(S) = Total number of outcomes

TERMS IN PROBABILITY

- Random Experiment
- Sample Sample
- Random variables
- Expected Value
- Independence
- Variance
- Mean

Random Experiment

An experiment whose result cannot be predicted, until it is noticed is called a random experiment. For example, when we throw a dice randomly, the result is uncertain to us. We can get any output between 1 to 6. Hence, this experiment is random.

Sample Space

A sample space is the set of all possible results or outcomes of a random experiment. Suppose, if we have thrown a dice, randomly, then the sample space for this experiment will be all possible outcomes of throwing a dice, such as;

Sample Space = { 1,2,3,4,5,6}

Random Variables

The variables which denote the possible outcomes of a random experiment are called random variables. They are of two types:

1. Discrete Random Variables
2. Continuous Random Variables

Discrete random variables take only those distinct values which are countable. Whereas continuous random variables could take an infinite number of possible values.

Independent Event

When the probability of occurrence of one event has no impact on the probability of another event, then both the events are termed as independent of each other. For example, if you flip a coin and at the same time you throw a dice, the probability of getting a 'head' is independent of the probability of getting a 6 in dice.

Mean

Mean of a random variable is the average of the random values of the possible outcomes of a random experiment. In simple terms, it is the expectation of the possible outcomes of the random experiment, repeated again and again or n number of times. It is also called the expectation of a random variable.

Expected Value

Expected value is the mean of a random variable. It is the assumed value which is considered for a random experiment. It is also called expectation, mathematical expectation or first moment. For example, if we roll a dice having six faces, then the expected value will be the average value of all the possible outcomes, i.e. 3.5.

Variance

Basically, the variance tells us how the values of the random variable are spread around the mean value. It specifies the distribution of the sample space across the mean.

CONCEPTS OF PROBABILITY

- Probability of an impossible event is phi or a null set.
- The maximum probability of an event is its sample space
- Probability of any event exists between 0 and 1.
- There cannot be a negative probability for an event.
- If A and B are two mutually exclusive outcomes

THEOREMS OF PROBABILITY : ADDITION

If two occurrences are denoted by the letters A and B, then the probability that at least one of the events will occur can be calculated as follows:

$$P(AUB) = P(A) + P(B) - P(AB).$$

THEOREMS OF PROBABILITY: MULTIPLICATION

The multiplication theorem on probability for dependent events can be extended for the independent events. From the theorem, we have, **$P(A \cap B) = P(A)\, P(B \mid A)$.** If the events A and B are independent, then, **$P(B \mid A) = P(B)$**. The above theorem reduces to
$P(A \cap B) = P(A)\, P(B)$.

BAYES THEOREM (THEOREM WITHOUT PROOF)

Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes.

$$P(A/B) = \frac{P(B/A)\, P(A)}{P(B)}$$

DISCRETE RANDOM SAMPLING DEFINITION

A random variable is called discrete. if it has either a finite or a countable number of possible values. A random variable is called continuous. if its possible values contain a whole interval of numbers.

MEAN OF DISCRETE RANDOM SAMPLING

$$E[X] = \sum x\, P(X=x\ )$$

VARIANCE OF DISCRETE RANDOM SAMPLING

$$Var[X] = \sum (x-\mu)2\, P(X=x)$$

THEORETICAL DISTRIBUTIONS

BINOMIAL DISTRIBUTION

The prefix ‘Bi’ means two or twice. A binomial distribution can be understood as the probability of a trail with two and only two outcomes. It is a type of distribution that has two different outcomes namely, ‘success’ and ‘failure’. Also, it is applicable to discrete random variables only.

POISSON DISTRIBUTION :

The Poisson Distribution is a theoretical discrete probability distribution that is very useful in situations where the events occur in a continuous manner. Poisson Distribution is utilized to determine the probability of exactly x_0 number of successes taking place in unit time. Let us now discuss the Poisson Model.

NORMAL DISTRIBUTION :

The Normal Distribution defines a probability density function f(x) for the continuous random variable X considered in the system. The random variables which follow the normal distribution are ones whose values can assume any known value in a given range.

PROPERTIES OF BINOMIAL DISTRIBUTION

1. There are two possible outcomes: true or false, success or failure, yes or no.
2. There is ‘n’ number of independent trials or a fixed number of n times repeated trials.
3. The probability of success or failure remains the same for each trial.
4. Only the number of success is calculated out of n independent trials.
5. Every trial is an independent trial, which means the outcome of one trial does not affect the outcome of another trial.

PROPERTIES OF POISSON MODEL :

1. The event or success is something that can be counted in whole numbers.
2. The probability of having success in a time interval is independent of any of its previous occurrence.
3. The average frequency of successes in a unit time interval is known.
4. The probability of more than one success in unit time is very low.

PROPERTIES OF NORMAL DISTRIBUTION :

Its shape is symmetric.

1. The mean and median are the same and lie in the middle of the distribution
2. Its standard deviation measures the distance on the distribution from the mean to the inflection point (the place where the curve changes from an “upside-down-bowl” shape to a “right-side-up-bowl” shape).
3. Because of its unique bell shape, probabilities for the normal distribution follow the Empirical Rule, which says the following:

4. About 68 percent of its values lie within one standard deviation of the mean. To find this range, take the value of the standard deviation, then find the mean plus this amount, and the mean minus this amount.
5. About 95 percent of its values lie within two standard deviations of the mean.
6. Almost all of its values lie within three standard deviations of the mean.

USES OF BINOMIAL DISTRIBUTION

1. Each replication of the process results in one of two possible outcomes (success or failure),
2. The probability of success is the same for each replication, and
3. The replications are independent, meaning here that a success in one patient does not influence the probability of success in another.

USES OF POISSON DISTRIBUTION

1. Individual events happen at random and independently. That is, the probability of one event doesn’t affect the probability of another event.
2. You know the mean number of events occurring within a given interval of time or space. This number is called λ (lambda), and it is assumed to be constant.

USES AND APPLICATION OF NORMAL DISTRIBUTION

1. It’s application goes beyond describing distributions
2. It is used by researchers and modelers.
3. The major use of normal distribution is the role it plays in statistical inference.
4. The z score along with the t –score, chi-square and F-statistics is important in hypothesis testing.
5. It helps managers/management make decisions.

APPLICATION OF BINOMIAL DISTRIBUTION

1. Manufacturing company uses binomial distribution to detect the defective goods or items
2. In the clinical trial binomial trial is used to detect the effectiveness of the drug
3. Moreover binomial trial is used in various field such as market research

APPLICATION OF POISSON DISTRIBUTION

1. The count of - particles emitted per unit of time is useful in analysis of any radio-active substance.
2. Number of telephone calls received at a given switch board per small unit of time.
3. Number of deaths per day or week due to a rare disease in a big hospital
4. In industrial production to find the proportion of defects per unit length, per unit area etc.
5. The count of bacteria per c.c. in blood
6. Distribution of number of mis-prints per page of a book

UNIT - IV

SAMPLING MEANING

A sample is a subset of individuals from a larger population. Sampling means selecting the group that you will actually collect data from in your research. For example, if you are researching the opinions of students in your university, you could survey a sample of 100 students.

THEORITICAL BASIS OF SAMPLING

METHODS OF SAMPLING

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are::

- Probability Sampling
- Non-probability Sampling

What is Probability Sampling?

The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of the population.

PROBABILITY SAMPLING TYPES

Probability Sampling methods are further classified into different types, such as simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Let us discuss the different types of probability sampling methods along with illustrative examples here in detail.

SIMPLE RANDOM SAMPLING

In simple random sampling technique, every item in the population has an equal and likely chance of being selected in the sample. Since the item selection entirely depends on the chance, this method is known as “Method of chance Selection”. As the sample size is large, and the item is chosen randomly, it is known as “Representative Sampling”.

Example:

Suppose we want to select a simple random sample of 200 students from a school. Here, we can assign a number to every student in the school database from 1 to 500 and use a random number generator to select a sample of 200 numbers.

SYSTEMATIC SAMPLING

In the systematic sampling method, the items are selected from the target population by selecting the random selection point and selecting the other methods after a fixed sample interval. It is calculated by dividing the total population size by the desired population size.

Example:

Suppose the names of 300 students of a school are sorted in the reverse alphabetical order. To select a sample in a systematic sampling method, we have to choose some 15 students by randomly selecting a starting number, say 5. From number 5 onwards, will select every 15th person from the sorted list. Finally, we can end up with a sample of some students.

STRATIFIED SAMPLING

In a stratified sampling method, the total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group, the statisticians randomly select the sample.

For example :

there are three bags (A, B and C), each with different balls. Bag A has 50 balls, bag B has 100 balls, and bag C has 200 balls. We have to choose a sample of balls from each bag proportionally. Suppose 5 balls from bag A, 10 balls from bag B and 20 balls from bag C.

CLUSTERED SAMPLING

In the clustered sampling method, the cluster or group of people are formed from the population set. The group has similar significatory characteristics. Also, they have an equal chance of being a part of the sample. This method uses simple random sampling for the cluster of population.

Example:

An educational institution has ten branches across the country with almost the number of students. If we want to collect some data regarding facilities and other things, we can’t travel to every unit to collect the required data. Hence, we can use random sampling to select three or four branches as clusters.

NON PROBABILITY SAMPLING METHODS

What is Non-Probability Sampling?

The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.

NON-PROBABILITY SAMPLING TYPES

Non-probability Sampling methods are further classified into different types, such as convenience sampling, consecutive sampling, quota sampling, judgmental sampling, snowball sampling.

CONVENIENCE SAMPLING

In a convenience sampling method, the samples are selected from the population directly because they are conveniently available for the researcher. The samples are easy to select, and the researcher did not choose the sample that outlines the entire population.

Example:

In researching customer support services in a particular region, we ask your few customers to complete a survey on the products after the purchase. This is a convenient way to collect data. Still, as we only surveyed customers taking the same product. At the same time, the sample is not representative of all the customers in that area.

CONSECUTIVE SAMPLING

Consecutive sampling is similar to convenience sampling with a slight variation. The researcher picks a single person or a group of people for sampling. Then the researcher researches for a period of time to analyze the result and move to another group if needed.

QUOTA SAMPLING

In the quota sampling method, the researcher forms a sample that involves the individuals to represent the population based on specific traits or qualities. The researcher chooses the sample subsets that bring the useful collection of data that generalizes the entire population.

PURPOSIVE OR JUDGMENTAL SAMPLING

In purposive sampling, the samples are selected only based on the researcher's knowledge. As their knowledge is instrumental in creating the samples, there are the chances of obtaining highly accurate answers with a minimum marginal error. It is also known as judgmental sampling or authoritative sampling.

SNOWBALL SAMPLING

Snowball sampling is also known as a chain-referral sampling technique. In this method, the samples have traits that are difficult to find. So, each identified member of a population is asked to find the other sampling units. Those sampling units also belong to the same targeted population.

PROBABILITY SAMPLING VS NON-PROBABILITY SAMPLING METHODS

The below table shows a few differences between probability sampling methods and non-probability sampling methods.

CENSUS METHOD

A census method is that process of the statistical list where all members of a population are analysed. The population relates to the set of all observations under concern. For instance, if you want to carry out a study to find out student's feedback about the amenities of your school, then all the students of your school would form a component of the 'population' for your study.

SAMPLING METHOD

A sampling method is a process for choosing sample members from a population. Three (3) common sampling methods are:

- Simple random sampling
- Stratified sampling
- Cluster sampling

RANDOM SAMPLING MEANING

Random sampling is referred to as that sampling technique where the probability of choosing each sample is equal.

FORMULA : $P = 1-(1-(1/N))n$

DEFINITION

Random sampling is a method of choosing a sample of observations from a population to make assumptions about the population. It is also called probability sampling

ADVANTAGES

- Easy to implement.
- Each member of the population has an equal chance of being chosen.
- Free from bias.

DISADVANTAGES

- If the sampling frame is large random sampling may be impractical.
- A complete list of the population may not be available.
- Minority subgroups within the population may not be present in sample.

SAMPLE SUM USING FORMULA

$$P = n/N = 100/1000 = 10\%$$

And the chance of selection of an employee more than once is;

$P = 1-(1-(1/N))n$

$P = 1 – (999/1000)100$

$P = 0.952$

$P \approx 9.5\%$

NON - RANDOM SAMPLING MEANING

Non-random sampling is a sampling technique where the sample selection is based on factors other than just random chance. In other words, non-random sampling is biased in nature.

DEFINITION

Non-Random sampling: In a non-random sampling method, all the units of the population do not have an equal chance of being selected and convenience or judgement of the investigator plays an important role in the selection of the sample. They are mainly selected on the basis of judgement, purpose, convenience or quota.

ADVANTAGES

- Quick and convenient
- As a general rule, non-probability samples can be constituted quickly, which allows the survey to be launched, executed and finished in shorter times.
- Inexpensive
- It usually only takes a few hours to an interviewer to conduct such a survey. As well, non-probability samples are generally not spread out geographically, therefore travelling expenses for interviewers are low. In web panels or crowdsourcing, no interviewers are necessary. Tracing and persuasion of non-respondents are not required or less demanding.
- Reduce respondent burden
- In the case of volunteer sampling or crowdsourcing, respondents volunteer to participate in the survey without being solicited personally.

DISADVANTAGES

- Selection bias
- In order to make inferences about the population, it requires strong assumptions about the similarity between the sample and the population even though the respondents are self-selected. Due to the selection bias presented in all non-probability samples, these are often dangerous assumptions to make. When generalization to the whole population is to be made, probability sampling should be performed instead.
- Noncoverage (undercoverage) bias
- Since some units in the population can have no chance of being included in the sample, it results noncoverage bias. For example, people without the internet at home might never be selected for a web panel and may differ from those with the internet.
- Difficulty of assessing the quality
- It is impossible to determine the probability that a unit in the population is selected for the sample, so reliable estimates and estimates of sampling error cannot be computed.

MERITS OF SAMPLING

(1) This method is free from personal bias of the investigator.

(2) Each and every item of the universe stands equal chances of being selected.
(3) The universe gets fairly represented by the sample.
(4) This is a very simple and straightforward method.

ADVANTAGES OF SAMPLING

- Sampling saves time to a great extent by reducing the volume of data. You do not go through each of the individual items.
- Sampling Avoids monotony in works. You do not have to repeat the query again and again to all the individual data.
- When you have limited time, survey without using sampling becomes impossible. It allows us to get near-accurate results in much lesser time
- When you use proper methods, you are likely to achieve higher level of accuracy by using sampling than without using sampling in some cases due to reduction in monotony, data handling issues etc.
- By using sampling, you can get detailed information on the data even by employing small amount of resources.

DISADVANTAGES OF SAMPLING

- Since choice of sampling method is a judgmental task, there exist chances of biasness as per the mindset of the person who chooses it.
- Improper selection of sampling techniques may cause the whole process to defunct.
- Selection of proper size of samples is a difficult job.
- Sampling may exclude some data that might not be homogenous to the data that are taken. This affects the level of accuracy in the results.

SAMPLING ERROR MEANING

Sampling error can be measured in different ways, but in reality, the error obtained is almost always an estimate of the actual error rather than the absolute measure of the error. To calculate any true population, first, we have to calculate the sample value.

DEFINITION

Sampling error is defined as the amount of inaccuracy in estimating some value, which occurs due to considering a small section of the population, called the sample, instead of the whole population. It is also called an error.

SAMPLING ERROR FORMULA

$$\textbf{Sampling Error, S. E} = (1/\sqrt{N})\ 100$$

METHODS TO REDUCE SAMPLING ERRORS

- Increasing sample size

- Stratification

ADVANTAGES OF SAMPLING ERROR

Sampling error refers to the difference between a sample statistic and the corresponding population parameter. It is an inherent aspect of sampling and cannot be entirely eliminated.

ADVANTAGES OF SAMPLING ERROR:

- It allows for estimation of the level of uncertainty in a sample.
- It can be used to construct confidence intervals, which provide a range of plausible values for a population parameter based on the sample data.
- It provides a way to test hypotheses about population parameters using sample data.
- Sampling error can be reduced by using a larger sample size.
- In general, sampling error is a useful tool for understanding the uncertainty and generalizability of results from sample data.

DISADVANTAGES OF SAMPLING ERROR

- It can lead to inaccurate or unreliable estimates of population parameters.
- It can result in biased or misleading conclusions about a population, if the sample is not representative of the population.
- It can be reduced by increasing the sample size, but this is not always possible or practical.
- It can be affected by factors such as nonresponse and measurement error, which can further complicate the interpretation of the results.
- It can lead to a lack of precision in the estimates, making it difficult to make comparisons or conclusions with a high level of confidence.

NON SAMPLING MEANING

Non-sampling error refers to all sources of error that are unrelated to sampling. Non-sampling errors are present in all types of survey, including censuses and administrative data.

DEFINITION

In statistics, non-sampling error is a catch-all term for the deviations of estimates from their true values that are not a function of the sample chosen, including various systematic errors and random errors that are not due to sampling. Non-sampling errors are much harder to quantify than sampling errors.

ADVANTAGES OF NON SAMPLING ERROR

- They can often be identified and corrected, unlike sampling errors which are inherent to the sampling process.

- Non-sampling errors can be controlled by careful design and implementation of the survey.
- They can be estimated, which allows for adjustments to be made to the survey results to account for their impact.
- Non-sampling errors may be less frequent and less severe than sampling errors.
- Because non-sampling errors do not stem from random processes, they can often be attributed to specific causes, which allows for targeted interventions to reduce them.

DISADVANTAGES OF NON SAMPLING ERROR

- Data entry errors: mistakes made when entering data into a computer or other database
- Measurement errors: inaccuracies in the way data is collected, such as using a faulty instrument or not properly training survey administrators
- Nonresponse bias: when certain groups of people do not participate in a survey or study, leading to a biased sample
- Response bias: when participants give inaccurate or untruthful answers, due to social desirability bias, leading to inaccurate results

UNIT- V

TESTING OF HYPOTHESIS

HYPOTHESIS TESTING MEANING

Hypothesis testing is a systematic procedure for deciding whether the results of a research study support a particular theory which applies to a population. Hypothesis testing uses sample data to evaluate a hypothesis about a population.

DEFINITION

Hypothesis testing can be defined as a statistical tool that is used to identify if the results of an experiment are meaningful or not. It involves setting up a null hypothesis and an alternative hypothesis. These two hypotheses will always be mutually exclusive.

NULL HYPOTHESIS

The null hypothesis is a concise mathematical statement that is used to indicate that there is no difference between two possibilities. In other words, there is no difference between certain characteristics of data. This hypothesis assumes that the outcomes of an experiment are based on chance alone. It is denoted as H_O

Hypothesis testing is used to conclude if the null hypothesis can be rejected or not. Suppose an experiment is conducted to check if girls are shorter than boys at the age of 5. The null hypothesis will say that they are the same height.

ALTERNATIVE HYPOTHESIS

The alternative hypothesis is an alternative to the null hypothesis. It is used to show that the observations of an experiment are due to some real effect. It indicates that there is a statistical significance between two possible outcomes and can be denoted as H_1 or H_A

For the above-mentioned example, the alternative hypothesis would be that girls are shorter than boys at the age of 5.

TYPE I ERROR

A type I error appears when the null hypothesis (H_O) of an experiment is true, but still, it is rejected. It is stating something which is not present or a false hit. A type I error is often called a false positive

TYPE II ERROR

A type II error appears when the null hypothesis is false but mistakenly fails to be refused. It is losing to state what is present and a miss. A type II error is also known as false negative

HYPOTHESIS TESTING P VALUE

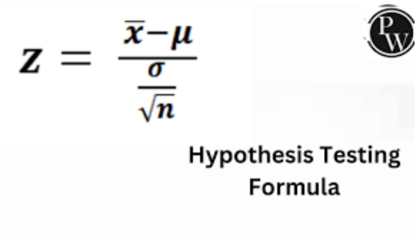
In hypothesis testing, the p value is used to indicate whether the results obtained after conducting a test are statistically significant or not. It also indicates the probability of making an error in rejecting or not rejecting the null hypothesis.This value is always a number between 0 and 1. The p value is compared to an alpha level, α or significance level. The alpha level can be defined as the acceptable risk of incorrectly rejecting the null hypothesis. The alpha level is usually chosen between 1% to 5%.

HYPOTHESIS TESTING CRITICAL REGION

All sets of values that lead to rejecting the null hypothesis lie in the critical region. Furthermore, the value that separates the critical region from the non-critical region is known as the critical value.

HYPOTHESIS TESTING FORMULA

Depending upon the type of data available and the size, different types of hypothesis testing are used to determine whether the null hypothesis can be rejected or not.



$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

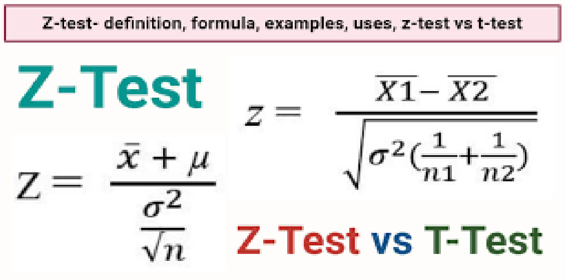
Hypothesis Testing Formula

TYPES OF HYPOTHESIS TESTING

Selecting the correct test for performing hypothesis testing can be confusing. These tests are used to determine a test statistic on the basis of which the null hypothesis can either be rejected or not rejected. Some of the important tests used for hypothesis testing are given below.

HYPOTHESIS TESTING Z TEST

A z test is a way of hypothesis testing that is used for a large sample size (n ≥ 30). It is used to determine whether there is a difference between the population mean and the sample mean when the population standard deviation is known. It can also be used to compare the mean of two samples. It is used to compute the z test statistic. The formulas are given as follows:



Z-test- definition, formula, examples, uses, z-test vs t-test

Z-Test

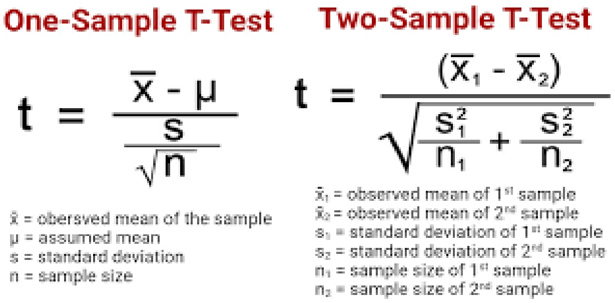
$$Z = \frac{\bar{x} + \mu}{\frac{\sigma^2}{\sqrt{n}}}$$

$$z = \frac{\overline{X1} - \overline{X2}}{\sqrt{\sigma^2(\frac{1}{n1} + \frac{1}{n2})}}$$

Z-Test vs T-Test

HYPOTHESIS TESTING T TEST

The t test is another method of hypothesis testing that is used for a small sample size (n < 30). It is also used to compare the sample mean and population mean. However, the population standard deviation is not known. Instead, the sample standard deviation is known. The mean of two samples can also be compared using the t test.



One-Sample T-Test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$\bar{x}$ = obersved mean of the sample
μ = assumed mean
s = standard deviation
n = sample size

Two-Sample T-Test

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ = observed mean of 1st sample
$\bar{x}_2$ = observed mean of 2nd sample
s_1 = standard deviation of 1st sample
s_2 = standard deviation of 2nd sample
n_1 = sample size of 1st sample
n_2 = sample size of 2nd sample

HYPOTHESIS TESTING CHI SQUARE

The Chi square test is a hypothesis testing method that is used to check whether the variables in a population are independent or not. It is used when the test statistic is chi-squared distributed.

$$x_c^2 = \frac{\Sigma\ (O_i\ - E_i)^2}{E_i}$$

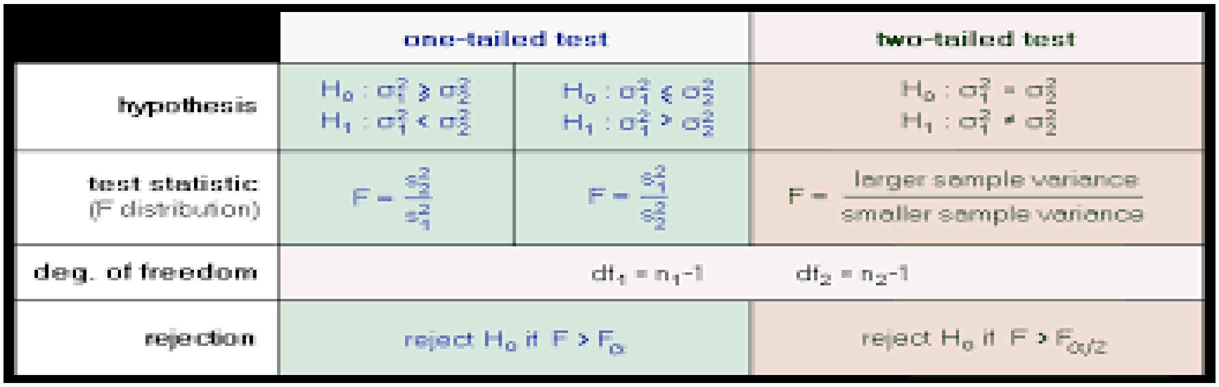
F TEST MEANING

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis.

DEFINITION

F test can be defined as a test that uses the f test statistic to check whether the variances of two samples (or populations) are equal to the same value. To conduct an f test, the population should follow an f distribution and the samples must be independent events. On conducting the hypothesis test, if the results of the f test are statistically significant then the null hypothesis can be rejected otherwise it cannot be rejected.

FORMULA :



	one-tailed test		two-tailed test
hypothesis	$H_0 : \sigma_1^2 \geqslant \sigma_2^2$ $H_1 : \sigma_1^2 < \sigma_2^2$	$H_0 : \sigma_1^2 \leqslant \sigma_2^2$ $H_1 : \sigma_1^2 > \sigma_2^2$	$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$
test statistic (F distribution)	$F = \frac{s_2^2}{s_1^2}$	$F = \frac{s_1^2}{s_2^2}$	$F = \frac{\text{larger sample variance}}{\text{smaller sample variance}}$
deg. of freedom	$df_1 = n_1 - 1$		$df_2 = n_2 - 1$
rejection	reject H_0 if $F > F_\alpha$		reject H_0 if $F > F_{\alpha/2}$