BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Financial Statistics

Course Code: 23ST09DEC

Unit-I

Probability in Financial Statistics

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. S. Soundarya
Guest Faculty
Department of Statistics

Introduction

Financial statistics consist of a comprehensive set of stock and flow data on the financial assets and liabilities of all sectors of an economy. The financial statistics are organized and presented in formats designed to show financial flows among the sectors of an economy and corresponding financial asset and liability positions.

Define Financial Statistics

Financial statistics are a comprehensive set of data on stocks and flows of financial assets and liabilities between all sectors of an economy and with the rest of the world, on a from-whom-to-whom basis where applicable.

The use financial statistics

There are six basic ratios that are often used to pick stocks for investment portfolios. Ratios include the working capital ratio, the quick ratio, and earnings per share (EPS), price-earnings (P/E), debt-to-equity, and return on equity (ROE).

Working Capital Ratio:

This ratio measures a company's short-term liquidity by comparing its current assets to its current liabilities. A higher ratio indicates the company's ability to meet its short-term obligations.

Quick Ratio:

Also known as the acid-test ratio, it measures a company's ability to meet its short-term liabilities with its most liquid assets. It excludes inventory from current assets as inventory might not be quickly converted to cash.

Earnings Per Share (EPS):

This ratio indicates the portion of a company's profit allocated to each outstanding share of common stock. It's a crucial metric for assessing a company's profitability and potential for future growth.

Price-Earnings (P/E) Ratio:

This ratio compares a company's stock price to its earnings per share. It helps investors gauge the market's valuation of the company's earnings and future growth potential.

Debt-to-Equity Ratio:

This ratio compares a company's total debt to its shareholder equity. It indicates the extent to which a company is financed by debt and provides insights into its financial risk and stability.

Return on Equity (ROE):

This ratio measures a company's profitability relative to its shareholder equity. It reflects how effectively a company is using its equity to generate profits.

Flow accounts

Flow accounts, also known as financial accounts, are a part of national accounts that track changes in economic activity over time. They are statements that record the sources and uses of funds, including:

- Savings: The difference between current income and current expenditure
- Depreciation: The allocation of the cost of an asset over its useful life
- Non-financial capital acquisition: The acquisition of non-financial assets
- Financial instrument transactions: Transactions involving financial assets like currency, loans, debt securities, equity, and more

Flow accounts are an extension of the Income and Expenditure Accounts (IEA), which measure aggregate economic activity. They are integrated with stocks, which are a unit's holdings of assets and liabilities at a specific time.

Flow accounts are used to create macro-financial indicators that help analyze economic sectors and compare countries. These indicators include: financial net wealth, indebtedness, corporate leverage, household net lending, and financial intermediation ratio.

stock account

This is an account one holds with a stock broker who is a member of a recognized stock exchange. This account is required to carry out buy or sell transactions in the securities markets. A stock broker, being a member of the stock exchange, is registered with and regulated by the exchange.

Accounting is the discipline of calculating, processing, and communicating financial information for businesses and individuals. Stock accounting is the type of accounting that covers these financial operations and responsibilities of the business's stock, accurately depicting the assets of the company.

Equities – If there is an investment of the account owner or common stocks, retained earnings then these will fall under equities.

Two-dimensional financial statistics

Two-dimensional financial statistics do not provide information on counterpart sectors. For example, it can show the purchase by the nonfinancial corporations (NFCs) sector of debt securities but not identify the sector that issued these securities.

Two-dimensional financial statistics can be represented in a financial balance sheet, which is a table that shows the closing stocks of nonfinancial corporations and general government. The table may include the following columns: Closing stocks, Currency and deposits, Debt securities, and Loans.

Here's an example of a two-dimensional financial balance sheet:

Other financial statistics include:

- Production
- Value added/GDP
- Distribution of income
- Disposable income
- Use of income/consumption
- Stocks
- Saving
- Other flows
- Nonfinancial assets
- Capital

A company's financial performance can be evaluated externally using financial statements, which include the balance sheet, income statement, and statement of cash flow.

Three-dimensional financial statistics

Three-dimensional financial statistics can refer to the three dimensions of a financial market, or to the three dimensions of financial inclusion:

Financial market dimensions

The three dimensions of a financial market are price, quantity, and time. A three-dimensional supply-demand-price-disequilibrium chart can be used to model a financial market over time.

Financial inclusion dimensions

The three dimensions of financial inclusion are access to financial services, usage of financial services, and the quality of the products and services.

Investment returns

Investment returns are not distributed like natural phenomena, so standard deviation is only a good rule of thumb. The differences in how heights and returns are distributed can lead to misunderstandings when estimating the probability and severity of downside risk.

Analytics

Analytics has three dimensions. Some people think of analytics as predictive capabilities, advanced math and statistics, speed of thought visualization, or multi-dimensional analysis.

Financial market: three-dimensional (price, quantity, time) supply-demand-price-disequilibrium chart-over time.

Scope of Financial Statistics

1. **Planning**

   The financial manager projects how much money the company will need in order to maintain positive cash flow, allocate funds to grow or add new products or services and cope with unexpected events, and shares that information with business colleagues.

2. **Budgeting**

   The financial manager allocates the company's available funds to meet costs, such as mortgages or rents, salaries, raw materials, employee T&E and other obligations. Ideally there will be some left to put aside for emergencies and to fund new business opportunities.

   Companies generally have a master budget and may have separate sub documents covering, for example, cash flow and operations; budgets may be static or flexible.

3. **Managing and assessing risk**

   Line-of-business executives look to their financial managers to assess and provide compensating controls for a variety of risks, including:

1. **Market risk**

Affects the business' investments as well as, for public companies, reporting and stock performance. May also reflect financial risk particular to the industry, such as a pandemic affecting restaurants or the shift of retail to a direct-to-consumer model.

2. **Credit risk**

The effects of, for example, customers not paying their invoices on time and thus the business not having funds to meet obligations, which may adversely affect creditworthiness and valuation, which dictates ability to borrow at favourable rates.

3. **Liquidity risk**

Finance teams must track current cash flow, estimate future cash needs and be prepared to free up working capital as needed.

4. **Operational risk**

This is a catch-all category and one new to some finance teams. It may include, for example, the risk of a cyber-attack and whether to purchase cyber security insurance, what disaster recovery and business continuity plans are in place and what crisis management practices are triggered if a senior executive is accused of fraud or misconduct.

4. Procedures

The financial manager sets procedures regarding how the finance team will process and distribute financial data, like invoices, payments and reports, with security and accuracy. These written procedures also outline who is responsible for making financial decisions at the company — and who signs off on those decisions.

Companies don't need to start from scratch; there are policy and procedure templates available for a variety of organization types, such as this one for nonprofits.

Random Variable

Defined over the sample space of a random experiment, is called a **random variable**. That is, the values of the random variable correspond to the outcomes of the random experiment.

Examples:

1. We throw two coins and count the number of heads.

2. We define X = 1 if the economy grows two consecutive quarters and X = 0, otherwise. (This is an example of a Bernouille (or indicator) RV.)

3. We read comments from IBM's CEO and compute IBM's return.

4. We count the days in a week the stock market has a positive return.

5. We look at a CEO and write his/her highest education degree

Basic Statistical Concepts for Finance

Basic Analysis

Depending on the role, some individuals may be using very advanced statistics for finance. However, many roles need to be able to carry out simple statistical analysis, such as taking an appropriate sample from a population and estimating parameters that explain the sample, such as the mean return or correlation.

Variables and Probability Distributions

The use of probability distributions can be found across all disciplines within the financial services industry. For example, the normal distribution can be found everywhere, from option valuation to measuring market risk, where a confidence interval approach can be used to calculate Value at Risk, or value at risk.

Just as businesses use statistical analysis to summarize data, finance analyzes variables like stock prices through probability distributions to help build financial models, calculate the expected value of a portfolio, and predict financial phenomena.

Time Series Analysis

Time series analysis, a fundamental tool in statistical finance, helps in studying trends over time, a vital aspect of financial prediction. Time series analysis is essential for examining historical trends and patterns in asset prices, aiding in the construction of portfolios that consider the time-dependent behaviours of financial instruments.

Regression Analysis

A multiple linear regression model is a statistical model used to help explain the relationship between independent variables and a dependent variable. Regression analysis can help economists predict how inflation, Gross Domestic Product(GDP), and other economic variables can impact stock returns. It can also be used to identify interest rate models.

Advanced Statistical Techniques for Finance

Stochastic Calculus Models

Advanced techniques used in financial engineering, such as financial mathematics and stochastic differential equations, are employed to price financial derivatives and manage risk effectively.

An example is the use of its formula to price financial derivatives and derive optimal hedging strategies, and to cover specific problems related to the derivation of interest rates used in interest rate models.

Nonlinear Time Series Analysis

Linear models assume a constant relationship, but in reality, the nonlinearity of financial markets demands sophisticated tools like nonlinear time series analysis for accurate predictions and risk assessments.

Balance Sheets

The balance sheets show stocks of nonfinancial and financial assets and liabilities on the date for which the balance sheet is compiled. The difference between total assets and total liabilities is net worth. For each group of assets and liabilities, and thus net worth, changes between the opening and closing balance sheets result from transactions and other flows recorded in the accumulation accounts.

Probability distribution

In Statistics, the probability distribution gives the possibility of each outcome of a random experiment or event. It provides the probabilities of different possible occurrences. Also read, events in probability, here.

To recall, the probability is a measure of uncertainty of various phenomena. Like, if you throw a dice, the possible outcomes of it, is defined by the probability. This distribution could be defined with any random experiments, whose outcome is not sure or could not be predicted.

Probability Distribution of Random Variables

A random variable has a probability distribution, which defines the probability of its unknown values. Random variables can be discrete (not constant) or continuous or both.

Probability distribution in finance

probability distributions are mathematical functions that calculate the likelihood of various outcomes. They are used to model and quantify financial interactions, such as option pricing and calculating a portfolio's Value-at-Risk (VaR).

Here are some ways that probability distributions are used in finance:

- **Anticipate returns**

Investors use probability distributions to predict future returns on assets, such as stocks.

- **Hedge risk**

Investors use probability distributions to calculate potential catastrophic events and hedge their risk.

- **Forecast demand**

Revenue analysts use probability distributions to model revenue probabilities for different scenarios and forecast demand.

- **Assess risk**

Risk analysts use probability distributions to analyze historical data and estimate the likelihood of future risks.

Application of Probability Distribution

Investors use probability distribution to predict returns overtime on assets, such as securities and to hedge their risk. Stock returns are often thought to be normally distributed. They show kurtosis with significant negative and positive returns. It appears to be more than a normal distribution would expect.

The distribution of stock returns has been defined as log-normal. It is because the stock prices are bounded by zero but give a possible unlimited upside. It shows up on a stock return plot with the distribution tails being larger in thickness.

Probability distributions are often commonly used in risk management to measure the probability. It also measures the sum of losses that an investment portfolio will experience based on a distribution of historical returns. One standard investment risk management metric is the Value-at-Risk (VaR). VaR yields the lowest loss that may occur, given a portfolio probability and time frame.

Terminology related to Probability

The terminology in probability listed below can help you comprehend probability ideas better.

- Experiment: An experiment is a trial or operation that is carried out to generate a specific result.
- Sample space: A sample space is the collection of all the probable outcomes of an experiment.. For example, a sample space of Tossing the coin is head and tail.
- Favourable Outcome: A favourable outcome is an event that has delivered the anticipated result or predicted event. When we roll two dice, for example, the possible/favourable outcomes of having the sum of the two dice as 4 are (1,3), (2,2), and (3,2). (3,1).
- Random Experiment: A random experiment is an experiment with a well-defined set of outcomes. When we toss a coin, for example, we know whether we will get heads or tails, but we don't know which one will appear.
- A random experiment's total number of results is referred to as an event.
- Equally Likely Events: Equally likely events are those that have the same possibility or probability of occurring. The outcome of one given event has no bearing to the outcome of the other. When we toss a coin, for example, we have an equal probability of getting a head or a tail.
- Exhaustive Events: The exhaustive event can be defined when the set of all possible outcomes of an experiment equals the sample space.

Types of Probability

Depending on the nature of the outcome or the method used to calculate the chance of an event occurring, several views or types of probabilities may exist. There are four major different types of probabilities:

- Classical Probability
- Axiomatic Probability
- Subjective Probability
- Empirical Probability

Let us understand each type of probability one by one,

Classical Probability

Classical probability which is also known as the theoretical probability which further states that if there are B equally likely outcomes in an experiment and event X has exactly A of them,

Then the probability of X is A/B, or P(X) = A/B.

It entails tossing a coin or rolling dice. It's computed by making a list of all the possible outcomes of the activity and keeping track of what actually happens. When throwing a coin, for example, the possible outcomes are heads or tails. If you toss the coin ten times, you must keep track of which outcome occurred each time.

Axiomatic Probability

A series of rules or axioms by Kolmogorov are applied to all kinds of axiomatic probability. The probability of each event occurring or not occurring can be calculated using these axioms, which are written as,

- The smallest and greatest probabilities are 0 and one, respectively.
- The chance of a specific happening is equal to one.
- Only one of two mutually exclusive events can occur at the same time, according to the union of events.

Subjective Probability

Subjective probability takes into account a person's personal belief on the probability of an event occurring. For example, a fan's opinion on the probability of a specific side winning a football match is based on their personal beliefs and feelings rather than a rigorous quantitative calculation.

Empirical Probability

Through thinking experiments, the empirical probability or experimental perspective evaluates probability. If a weighted die is rolled and we don't know which side has the weight, we can gain an idea of the probability of each outcome by rolling the die a certain number of times and counting the proportion of times the die produces that outcome, and therefore find the probability of that outcome.

Probability Terms and Definition

Some of the important probability terms are discussed here:

Term	Definition	Example
Sample Space	The set of all the possible outcomes to occur in any trial	1. Tossing a coin, Sample Space (S) = {H,T} 2. Rolling a die, Sample Space (S) = {1,2,3,4,5,6}
Sample Point	It is one of the possible results	In a deck of Cards: • 4 of hearts is a sample point. • The queen of clubs is a sample point.
Experiment or Trial	A series of actions where the outcomes are always uncertain.	The tossing of a coin, Selecting a card from a deck of cards, throwing a dice.
Event	It is a single outcome of an experiment.	Getting a Heads while tossing a coin is an event.
Outcome	Possible result of a trial/experiment	T (tail) is a possible outcome when a coin is tossed.
Complimentary event	The non-happening events. The complement of an event A is the event, not A (or A')	In a standard 52-card deck, A = Draw a heart, then A' = Don't draw a heart
Impossible Event	The event cannot happen	In tossing a coin, impossible to get both head and tail at the same time

Applications of Probability

Probability has a wide variety of applications in real life. Some of the common applications which we see in our everyday life while checking the results of the following events:

- Choosing a card from the deck of cards
- Flipping a coin
- Throwing a dice in the air

- Pulling a red ball out of a bucket of red and white balls
- Winning a lucky draw

Other Major Applications of Probability

- It is used for risk assessment and modelling in various industries
- Weather forecasting or prediction of weather changes
- Probability of a team winning in a sport based on players and strength of team
- In the share market, chances of getting the hike of share prices

Example: 1 Two dice are rolled, find the probability that the sum is:

1. **equal to 1**
2. **equal to 4**
3. **less than 13**

Solution:

To find the probability that the sum is equal to 1 we have to first determine the sample space S of two dice as shown below.

S = { (1,1),(1,2),(1,3),(1,4),(1,5),(1,6)

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

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

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

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

(6,1),(6,2),(6,3),(6,4),(6,5),(6,6) }

So, n(S) = 36

1) Let E be the event “sum equal to 1”. Since, there are no outcomes which where a sum is equal to 1, hence,

P(E) = n(E) / n(S) = 0 / 36 = 0

2) Let A be the event of getting the sum of numbers on dice equal to 4.

Three possible outcomes give a sum equal to 4 they are:

A = {(1,3),(2,2),(3,1)}

n(A) = 3

Hence, P(A) = n(A) / n(S) = 3 / 36 = 1 / 12

3) Let B be the event of getting the sum of numbers on dice is less than 13.

From the sample space, we can see all possible outcomes for the event B, which gives a sum less than B. Like:

(1,1) or (1,6) or (2,6) or (6,6).

So you can see the limit of an event to occur is when both dies have number 6, i.e. (6,6).

Thus, n(B) = 36

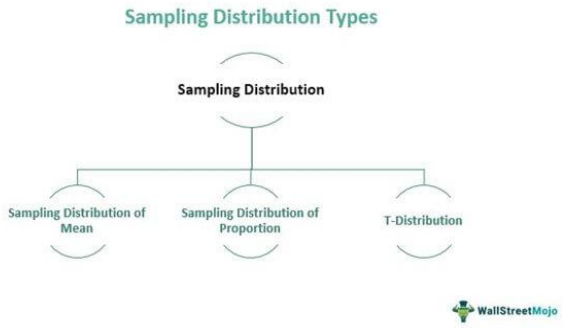
Hence,

P(B) = n(B) / n(S) = 36 / 36 = 1

Sampling distribution

A sampling distribution refers to a probability distribution of a statistic that comes from choosing random samples of a given population. Also known as a finite-sample distribution, it represents the distribution of frequencies on how spread apart various outcomes will be for a specific population.

Types of Sampling Distribution:



Sampling distribution of mean

The sampling distribution of the mean is a method for calculating the mean of each sample group from a population and plotting the data points. The graph of the sampling distribution of the mean is a normal distribution, with the mean of the sampling distribution representing the mean of the entire population.

Sampling distribution of proportion

A sampling distribution of a sample proportion is a distribution created by taking many random samples of a given size from a population and finding the proportion of a category of interest in each sample.

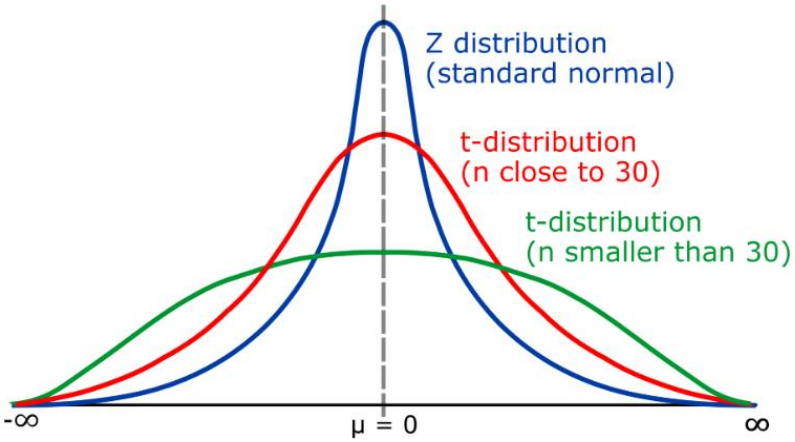
T-distribution

T-distribution is used when the sample size is very small or not much is known about the population. It is used to estimate the mean of the population, confidence intervals, statistical differences, and linear regression.

Student's t-distribution

The t-distribution describes the standardized distances of sample means to the population mean when the population standard deviation is not known, and the observations come from a normally distributed population. The t-distribution is used in hypothesis testing and confidence intervals for the mean of a single population, or the difference in means between two populations.

The Student's t-Distribution is a continuous probability distribution that is used to model the distribution of the t-statistic. It is commonly used in hypothesis testing when the sample size is small and the population variance is unknown. The Student's t-Distribution is denoted by t(df), where df denotes the degrees of freedom.



Applications of Student's t-Distribution:

Estimating population means

When the sample size is small or the population standard deviation is unknown, the t-distribution can be used to estimate the mean of a normally distributed population.

Hypothesis testing

The t-distribution can be used to test hypotheses about the mean of a normally distributed population, the difference between two means, or the coefficient of correlation.

Confidence intervals

The t-distribution can be used to construct confidence intervals for the mean of a normally distributed population.

t-tests

The t-distribution can be used to perform paired t-tests to compare the means of two related samples, or independent t-tests to compare the means of two independent samples.

Problem 1:

Imagine a company wants to test the claim that their batteries last more than 40 hours. Using a simple random sample of 15 batteries yielded a mean of 44.9 hours, with a standard deviation of 8.9 hours. Test this claim using a significance level of 0.05.

Hypotheses:

Null hypothesis (H_0): $\mu = 40$ and Alternative hypothesis (H_a): $\mu > 40$.

Given Data:

Sample mean ($\bar{x}$): 44.9, Population mean (μ): 40, Sample standard deviation (s): 8.9, Sample size (n): 15, and Degrees of freedom (df): n - 1 = 15 - 1 = 14.

Test Statistic:

$t = (\bar{x} - \mu) / (s / \sqrt{n}) = (44.9 - 40) / (8.9 / \sqrt{15}) = 2.13$

P-value: p-value = $P(t > 2.13) = 0.026$.

Conclusion:

Since the p-value (0.026) is less than the significance level ($\alpha = 0.05$), we reject the null hypothesis (H_0).

This means that there is sufficient evidence to support the alternative hypothesis that the population mean is greater than 40.

Problem 2:

The heights of six randomly chosen sailors are in inches: 63,65, 68, 69, 71 and 72. Those of 10 randomly chosen soldiers are 61,62,65,66,69,69,70,71,72 and 73. Discuss. the light that these data throw on the suggestion that sailors are on the'-average taller than soldiers.

Solution. If the heights of sailors and soldiers be represented by the variables X and Y respectively then the *Null Hypothesis* is, $H_0 : \mu_X = \mu_Y$, *i.e., the sailors are not on the average taller than the soldiers.*

Alternative Hypothesis, $H_1 : \mu_X > \mu_Y$ (Right-tailed).

Under H_0, the test statistic is :

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} = t_{14}$$

Sailors			*Soldiers*		
X	$d = X - A = X - 68$	d^2	Y	$D = Y - B = Y - 66$	D^2
63	−5	25	61	−5	25
65	−3	9	62	−4	16
68	0	0	65	−1	1
69	1	1	66	0	0
71	3	9	69	3	9
72	4	16	69	3	9
			70	4	16
			71	5	25
Total	0	60	72	6	36
			73	7	49
			Total	18	186

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2\right] = \frac{1}{14}(60 + 153{\cdot}6) = 15{\cdot}2571$$

$$\therefore \quad t = \frac{68 - 67{\cdot}8}{\sqrt{15{\cdot}2571}\left(\frac{1}{6} + \frac{1}{10}\right)^{1/2}} = \frac{0{\cdot}2}{\sqrt{15{\cdot}2571 \times 0{\cdot}2667}} = 0{\cdot}099$$

Tabulated $t_{0{\cdot}05}$ for 14 *d.f.* for single-tail test is 1·76.

Conclusion. Since calculated t is much less than 1·76, it is not at all significant at 5% levels of significance. Hence null hypothesis may be retained at 5% level of significance and we conclude that the data are inconsistent with the suggestion that the sailors are on the average taller than soldiers.

F-distribution

The F-distribution is used in hypothesis testing and confidence intervals for the variance ratio of two populations. In applied problems we may be interested in knowing whether the population variances are equal, based on the response of the random samples.

Applications of Student's F-Distribution:

Hypothesis testing

The F distribution is used to determine critical regions for hypothesis tests and to construct confidence intervals.

Comparing variances

The F distribution is used to compare the variances of two populations. For example, the F test can be used to determine if the variances of two normal populations are equal.

Analysis of variance (ANOVA)

The F distribution is used in ANOVA to compare the variation between groups and within groups. In ANOVA, the ratio of two mean squares (variances) is used to determine if the differences between sample means are statistically significant.

Regression analysis

The F distribution is used in regression analysis to test the overall significance of a regression model. This involves comparing the explained variance to the unexplained variance.

Example 14·21. *In one sample of 8 observations, the sum of the squares of deviations of the sample values from the sample mean was 84·4 and in the other sample of 10 observations it was 102·6. Test whether this difference is significant at 5 per cent level, given that the 5 per cent point of F for $n_1 = 7$ and $n_2 = 9$ degrees of freedom is 3·29.* **[Delhi Univ. B.Sc. (Maths Hons.), 1986]**

Solution. Here $n_1 = 8, n_2 = 10$

and $\Sigma(x - \bar{x})^2 = 84{\cdot}4, \quad \Sigma(y - \bar{y})^2 = 102{\cdot}6$

$$\therefore \quad S_X^2 = \frac{1}{n_1 - 1}\Sigma(x - \bar{x})^2 = \frac{84{\cdot}4}{7} = 12{\cdot}057$$

$$S_Y^2 = \frac{1}{n_2 - 1}\Sigma(y - \bar{y})^2 = \frac{102{\cdot}6}{9} = 11{\cdot}4$$

Under $H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$, *i.e.*, the estimates of σ^2 given by the samples are homogeneous, the test statistic is

$$F = \frac{S_X^2}{S_Y^2} = \frac{12{\cdot}057}{11{\cdot}4} = 1{\cdot}057$$

Tabulated $F_{0{\cdot}05}$ for (7, 9) d.f. is 3·29.

Since calculated $F < F_{0{\cdot}05}$, H_0 may be accepted at 5% level of significance.

Chi-Square distribution

A chi-square distribution is a continuous probability distribution. The shape of a chi-square distribution depends on its degrees of freedom, k. The mean of a chi-square distribution is equal to its degrees of freedom (k) and the variance is 2k.

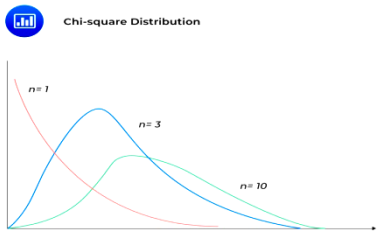
Applications of Student's Chi-square Distribution

Pearson's chi-square tests: A statistical test for categorical data that determines if data is significantly different from what was expected.

Goodness of fit test: Determines how well observed data represents the population.

Test for independence: Determines if there is a relationship between two categorical variables.

Finding confidence intervals: Used to estimate the population standard deviation of a normal distribution from a sample standard deviation.



Problem:

Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the table. Examine whether the nature of the area is related to voting preference in this election.

Votes for area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

E(620) = (1170 x 1000) / 2000 = 585,

E(380) = (830 x 1000) / 2000 = 415,

E(550) = (1170 x 1000) / 2000 = 585,

E(450) = (830 x 1000) / 2000 = 415

In a 2 x 2 contingency table, since d.f. = (2 - 1)(2 - 1) = 1,

$\chi^2 = \Sigma[(O - E)^2 / E] = (620 - 585)^2 / 585 + (380 - 415)^2 / 415 + (550 - 585)^2 / 585 + (450 - 415)^2 / 415$

$= (35)^2 [1 / 585 + 1 / 415 + 1 / 585 + 1 / 415]$

$= (1225) [2 \times 0.002409 + 2 \times 0.001709] = 10.0891$

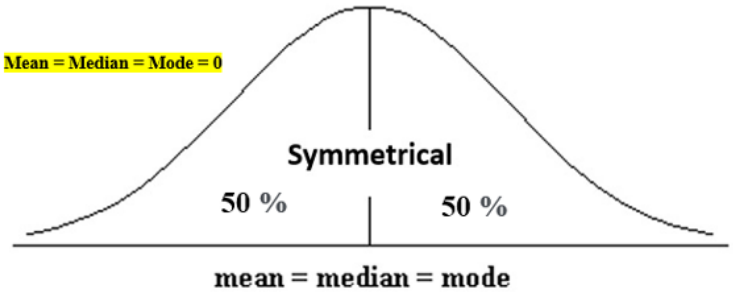
Tabulated χ^2 0.05 for (2 - 1)(2 - 1) = 1 d.f. is 3.841.

Conclusion

Since calculated χ^2 is much greater than the tabulated value, it is highly significant and null hypothesis is rejected at 5% level of significance. Thus we conclude that nature of area is related to voting preference in the election.

Skewness

Skewness is a statistical measure that assesses the asymmetry of a probability distribution. It quantifies the extent to which the data is skewed or shifted to one side.

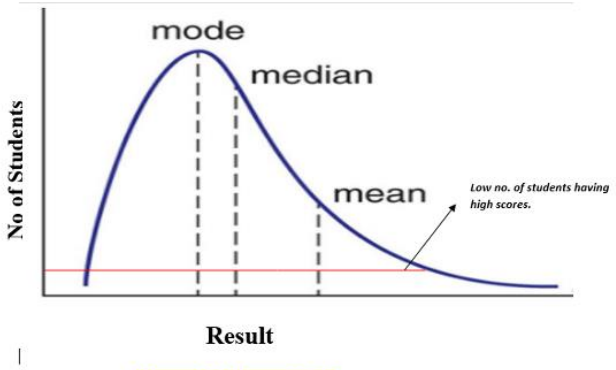


Types of Skewness

Positive Skewed or Right-Skewed (Positive Skewness)

In statistics, a positively skewed or right-skewed distribution has a long right tail. It is a sort of distribution where the measures are dispersing, unlike symmetrically distributed data

where all measures of the central tendency (mean, median, and mode) equal each other. This makes Positively Skewed Distribution a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.

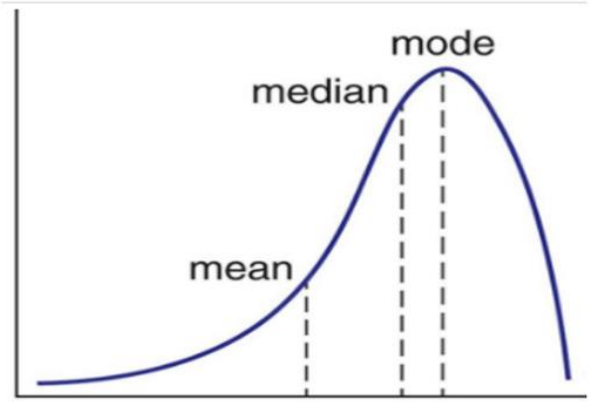


Mean > Median > Mode

In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the most frequent value.

Negative Skewed or Left-Skewed (Negative Skewness)

A distribution with a long left tail, known as negatively skewed or left-skewed, stands in complete contrast to a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.



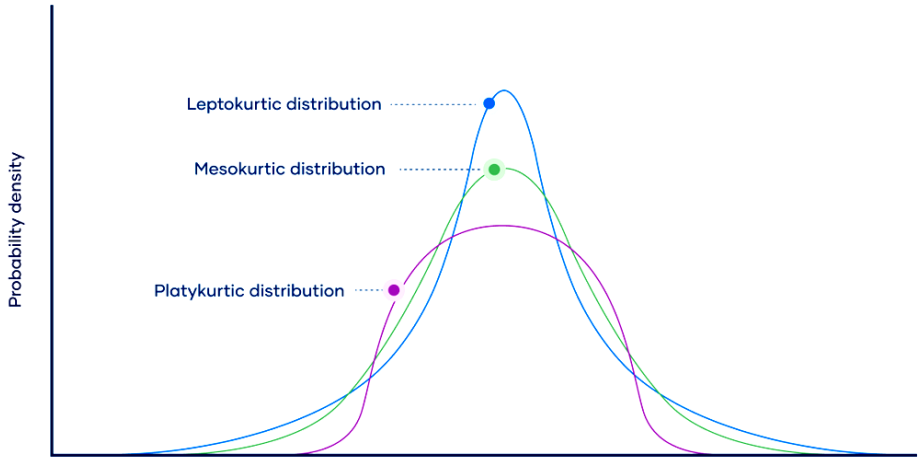
Mode > Median > mean

Kurtosis

Kurtosis is a statistical measure used to describe a characteristic of a dataset. When normally distributed data is plotted on a graph, it generally takes the form of a bell. This is called the bell curve.

Types of Excess Kurtosis

1. Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).
2. Mesokurtic (kurtosis same as the normal distribution).
3. Platykurtic or short-tailed distribution (kurtosis less than normal distribution).



Leptokurtic (Kurtosis > 3)

Leptokurtic has very long and thick tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. Extremely positive kurtosis indicates a distribution where more numbers are located in the tails of the distribution instead of around the mean.

Platykurtic (Kurtosis < 3)

Platykurtic having a thin tail and stretched around the center means most data points are present in high proximity to the mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

Mesokurtic (Kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.

Central limit Theorem

The **Central limit Theorem** states that when sample size tends to infinity, the sample mean will be normally distributed.

Statement

If X_i, (i = 1, 2, ..., n) be independent random variables such that $E(X_i) = \mu_i$ and $V(X_i) = \sigma^2$, then under certain very general conditions, the random variable $S_n = X_1 + X_2 + ... + X_n$, is asymptotically normal with mean and standard deviation where

$$\mu = \Sigma\mu_1 \text{ and } o^2 = \Sigma\, \sigma^2$$

S.No	Central limit theorem	Law of large numbers
1.	Sample mean will be normally distributed	Sample mean equals to population mean
2.	Can be executed for less data	Can be executed for very large numbers
3.	It will not converge to a number, it converges to a normal distribution.	Sample mean will converge to a Number
4.	Gives value and distribution of sample mean for n numbers	Gives value and distribution of sample mean for n numbers
5.	The standard deviation of the sample means equals the standard deviation of the population	Standard deviation declines as the size of population or sample increases

Law of Large Numbers (LLN):

The **Law of Large Number** states that when sample size tends to infinity, the sample mean equals to population mean.

A law of large numbers is a proposition given a set of conditions that are sufficient to guarantee the convergence of the sample mean $\bar{X}n$ to a constant as the sample size n increases.

Weak law of large numbers (WLLN)

Strong law of large numbers (SLLN)

Suppose X_1, X_2, ..., X_n are independent random variables with the same underlying distribution. In this case, we say that the X_i, are independent and identically-distributed, or i.i.d. In particular, the X_i, all have the same mean μ and standard deviation σ.

Let $\bar{X}n$ be the average of X_1,..., Xn

$$\bar{X}_n = X_1+X_2+ \ldots.+ X_n /n = \Sigma X_i/n$$

Note that $\bar{X}n$ is itself a random variable. The law of large numbers and central limit theorem tell us about the value and distribution of $\bar{X}n$, respectively.

1. Strong law of large numbers:

If the sequence $\{X_n\}$ converges almost surely, the law of large numbers is called a strong law of large numbers.

Almost everywhere convergence

A Sequence {fn} of functions defined on a set E is said to converge almost everywhere to f if

$$\lim_{n\to\infty} f_n(x) = f(x) \ \forall\, x \in E - E_1 \ where \ E_1 \subset E,$$

$$m(E_1) = 0.$$

2. Weak law of large numbers (WLLN):

If the sequence $\{X_n\}$ converges in probability, then the Law of large numbers is called a weak law of large numbers.

Convergence in Probability

A sequence of random variables X_1, X_2, X_3, ….. converges in probability to a random variable X, shown by $X_n \rightarrow X$, if

lim $P(|X_n - X| \geq €) = 0$, for all $€ > 0$.

Multivariate Normal Distributions

The multivariate normal distribution is a generalization of the univariate normal distribution to two or more variables. It is a distribution for random vectors of correlated variables, where each vector element has a univariate normal distribution. The joint density function as shown in the expression below:

$$\phi(\mathbf{x}) = \left(\tfrac{1}{2\pi}\right)^{p/2} |\Sigma|^{-1/2} \exp\{-\tfrac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)\}$$

Some things to note about the multivariate normal distribution:

1. The following term appearing inside the exponent of the multivariate normal distribution is a quadratic form:

$$(X-\mu)'\Sigma^{-1}(X-\mu)$$

This particular quadratic form is also called the squared Mahalanabis distance between the random vector X and the mean vector μ.

2. If the variables are uncorrelated then the variance covariance matrix will be a diagonal matrix with variances of the individual variables appearing on the main diagonal of the matrix and zeros everywhere else:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix}$$

Multivariate Normal Density Function:

In this case the multivariate normal density function simplifies to the expression below:

$$\phi(\mathbf{x}) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\{-\frac{1}{2\sigma_j^2}(x_j - \mu_j)^2\}$$

Covariance

Covariance is a statistical measure that indicates the direction of the linear relationship between two variables**.** It assesses how much two variables change together from their mean values.

Types of Covariance:

- **Positive Covariance**: When one variable increases, the other variable tends to increase as well, and vice versa.
- **Negative Covariance**: When one variable increases, the other variable tends to decrease.
- **Zero Covariance**: There is no linear relationship between the two variables; they move independently of each other.

Covariance is calculated by taking the average of the product of the deviations of each variable from their respective means. It is useful for understanding the direction of the relationship but not its strength, as its magnitude depends on the units of the variables.

It is an essential tool for understanding how variables change together and is widely used in various fields, including finance, economics, and science.

Covariance:

1. It is the relationship between a pair of random variables where a change in one variable causes a change in another variable.
2. It can take any value between – infinity to +infinity, where the negative value represents the negative relationship whereas a positive value represents the positive relationship.
3. It is used for the linear relationship between variables.
4. It gives the direction of relationship between variables.

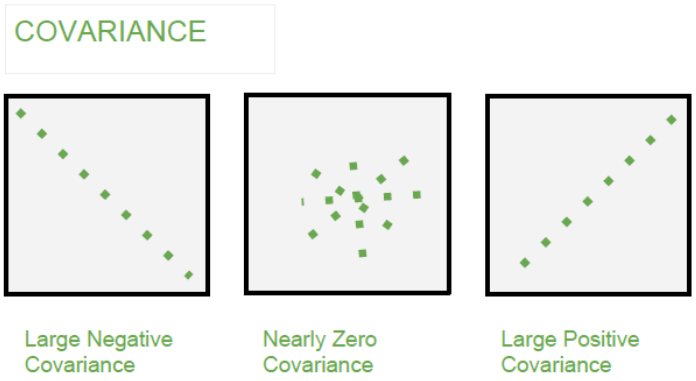
For Population:

$$Covri(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n}$$

For Sample:

$$Covari(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{n-1}$$

Here, x' and y' = mean of given sample set n = total no of sample xi and yi = individual sample of set



Correlation

Correlation is a **standardized measure of the strength and direction of the linear relationship between two variables**. It is derived from covariance and **ranges between -1 and 1**. Unlike covariance, which only indicates the direction of the relationship, correlation provides a standardized measure.

- **Positive Correlation (close to +1)**: As one variable increases, the other variable also tends to increase.

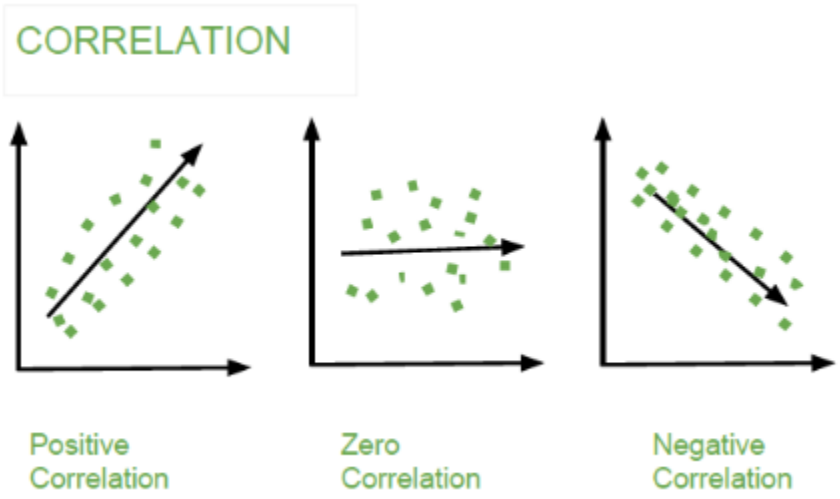
- **Negative Correlation (close to -1)**: As one variable increases, the other variable tends to decrease.
- **Zero Correlation**: There is no linear relationship between the variables.

The correlation coefficient ρ (rho) for variables X and Y is defined as:

1. It show whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. In this variable are indirectly related to each other.
4. It gives the direction and strength of relationship between variables.

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

Here, x' and y' = mean of given sample set n = total no of sample xi and yi = individual sample of set.



Difference between Covariance and Correlation

This table shows the **difference between Covariance and Covariance**:

Covariance	Correlation
Covariance is a measure of how much two random variables vary together	Correlation is a statistical measure that indicates how strongly two variables are related.

Covariance	Correlation
Involves the relationship between two variables or data sets	Involves the relationship between multiple variables as well
Lie between -infinity and +infinity	Lie between -1 and +1
Measure of correlation	Scaled version of covariance
Provides direction of relationship	Provides direction and strength of relationship
Dependent on scale of variable	Independent on scale of variable
Have dimensions	Dimensionless

Applications of Covariance and Correlation

Applications of Covariance

- **Portfolio Management in Finance**: Covariance is used to measure how different stocks or financial assets move together, aiding in portfolio diversification to minimize risk.
- **Genetics**: In genetics, covariance can help understand the relationship between different genetic traits and how they vary together.
- **Econometrics**: Covariance is employed to study the relationship between different economic indicators, such as the relationship between GDP growth and inflation rates.
- **Signal Processing**: Covariance is used to analyze and filter signals in various forms, including audio and image signals.
- **Environmental Science**: Covariance is applied to study relationships between environmental variables, such as temperature and humidity changes over time.

Applications of Correlation

- **Market Research**: Correlation is used to identify relationships between consumer behavior and sales trends, helping businesses make informed marketing decisions.
- **Medical Research**: Correlation helps in understanding the relationship between different health indicators, such as the correlation between blood pressure and cholesterol levels.
- **Weather Forecasting**: Correlation is used to analyze the relationship between various meteorological variables, such as temperature and humidity, to improve weather predictions.
- **Machine Learning**: Correlation analysis is used in feature selection to identify which variables have strong relationships with the target variable, improving model accuracy.

convert covariance to correlation

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Problem:

If the coefficient of correlation between x and y is 0.5 and their covariance is 16, and the SD of x is 4, then what is the SD of y?

Solution:

Given r = 0.5

Cov (x,y) = 16

$\sigma_x = 4$

$\sigma_y = \text{cov}(x,y) / r\sigma_x$

$= 16 / 0.5 \times 4$

$= 16 / (½) \times (42) = 162$

$= 8$