How and why ...

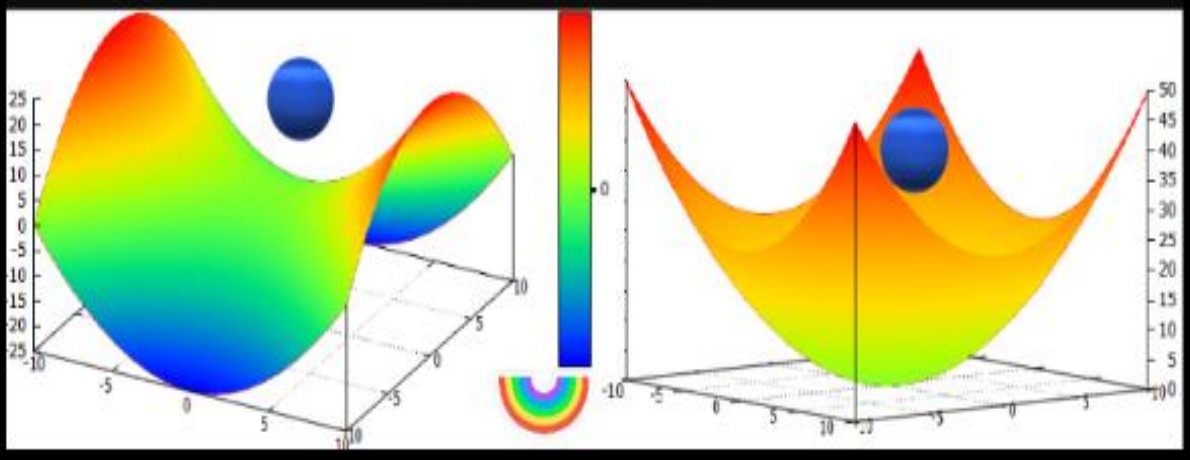
Statistics

Learn:

Exactly what you need to know about statistical ideas and techniques

Fundamental formulas and calculations

Core topics in scope of applications



Charles Cheng Xia, MD. MPH

Contents at a Glance

Examples at a Glance

Chi-squared Test:

Nonparametric Statistic Analysis:

Simple Linear Regression Equation and Correlation:

Multiple Linear Regression Equation and Correlation:

Table of Contents

Chapter 3: Probability Theory

Chapter 4:Statistical inference

Chapter 5: Analysis of Variance (ANOVA)

Chapter 6: Binomial Distribution

Chapter 7: Chi-squared Test

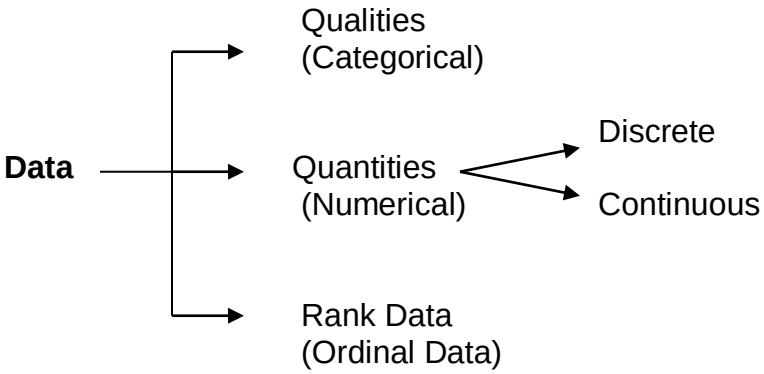
Chapter 8: Nonparametric Statistic Analysis

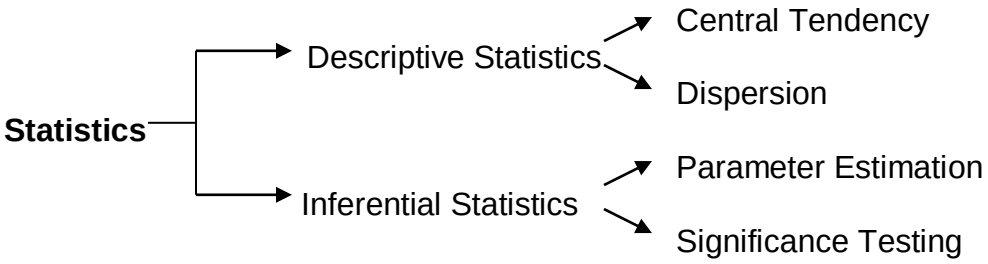
Chapter 9: Simple Regression Analysis

Chapter 10: Multiply Linear Regression (MLR)

Introduction

This book is focus on description of data and analysis of data with help you understand and learn exactly what you need to know about statistical ideas and techniques, fundamental formulas and calculations and statistical core topics in scope of applications. The book is mainly based on the following two illustration figures to extend the statistic contents.





For your better and/or easier understanding, this book includes more than forty examples in explanation and/or illustration, step by step, to let you understand or have ideas to understand on how and why the statistic formula and calculation be applied.

It is assumed that you've had a basic algebra background and can do some of the basic mathematical operations and understand some of basic notation used in algebra like x, y, summation sign, taking the square root, squaring a number, and so on.

About the Author
Charles Cheng Xia is a statistics and health educator with his Medical Doctor and Master of Public Health degree. His primary research interest lies in epidemiology, health economics and complex disease dynamics inferred from data science and mathematical modeling. Email: xc7788@gmail.com

Topics at a Glance

Chapter 1

Fundamental Concepts

What is Statistics?

Statistics is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data. A common aim of statistical analysis is to produce information about some chosen population.

Statistics include the four aspects:

1. Design: A proper statistical design is crucial for making sure the data that will enter the decision making process is of sufficient quality and thus the causal inference and/or estimation will be valid.

2. Collection of data: The collection of data is based on the statistics design to fulfill the statistics tasks with the research problem you explore informs; the type of data you'll collect; and the data collection method you'll use.

3. Sorting data: Data sorting is any process that involves arranging the data into some meaningful order to make it easier to understand, analyze or visualize.

4. Analysis of data: Statistical data analysis is a procedure of performing various statistical operations. It is a kind of quantitative research, which seeks to quantify the data, and typically, applies some form of statistical analysis. Quantitative data basically involves descriptive data, such as survey data and observational data.

This book is focus on description of data and analysis of data with help you understand and learn exactly what you need to know about statistical ideas and techniques, fundamental formulas and calculations and core topics in scope of applications.

Statistical Population

A population is a set of similar items or events which is of interest for some question or experiment. A statistical population can be a group of existing objects (e.g. the set of all stars within the Milky Way galaxy) or a hypothetical and potentially infinite group of objects conceived as a generalization from experience (e.g. the set of all possible hands in a game of poker).

Statistical Sample

In statistical inference, a subset of the population (a statistical sample) is chosen to represent the population in a statistical analysis. The ratio of the size of this statistical sample to the size of the population is called a sampling fraction. It is then possible to estimate the population parameters using the appropriate sample statistics.

In statistics and quantitative research methodology, a sample is a set of individuals or objects collected or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations.[citation needed] When conceived as a data set, a sample is often denoted by capital roman letters such X and Y, with its elements expressed in lower-case (e.g.,x3) and the sample size denoted by the letter n.

Typically, the population is very large, making a census or a complete enumeration of all the individuals in the population either impractical or impossible. The sample usually represents a subset of manageable size. Samples are collected and statistics are calculated from the samples, so that one can make inferences or extrapolations from the sample to the population.

The sample may be drawn from a population without replacement (i.e. no element can be selected more than once in the same sample), in which case it is a subset of a population; or with replacement (i.e. an element may appear multiple times in the one sample), in which case it is a multisubset.

Key points:

1. A selection that is chosen randomly (purely by chance, with no predictability).

2. Every member of the population being studied should have an equal chance of being selected.

Example: you want to survey 100 people at a football match about their main job. Asking just people in one area might give poor results as there may be a group of workmates there! So instead you can make a list using randomly chosen seat numbers from the whole stadium, then go and find each seat and interview the person there.

Mathematical Description of Random Sample

In mathematical terms, given a probability distribution F, a random sample of length n (where n may be any positive integer) is a set of realizations of n independent, identically distributed random variables with distribution F.

A sample concretely represents the results of n experiments in which the same quantity is measured. For example, if we want to estimate the average height of members of a particular population, we measure the heights of n individuals. Each measurement is drawn from the probability distribution F characterizing the population, so each measured height x_i is the realization of a random variable X_i with distribution F. Note that a set of random variables (i.e., a set of measurable functions) must not be confused with the realizations of these variables (which are the values that these random variables take).

Statistical Estimator

In statistics, an estimator is a rule for calculating an estimate of a given quantity based on observed data: thus the rule (the estimator), the quantity of interest (the estimand) and its result (the estimate) are distinguished.

Point and Interval Estimators

There are point and interval estimators. The point estimators yield single-valued results, although this includes the possibility of single vector-valued results and results that can be expressed as a single function. This is in contrast to an interval estimator, where the result would be a range of plausible values (or vectors or functions).

Estimation theory is concerned with the properties of estimators; that is, with defining properties that can be used to compare different estimators (different rules for creating estimates) for the same quantity, based on the same data. Such properties can be used to determine the best rules to use under given circumstances.

However, in robust statistics, statistical theory goes on to consider the balance between having good properties, if tightly defined assumptions hold, and having less good properties that hold under wider conditions.

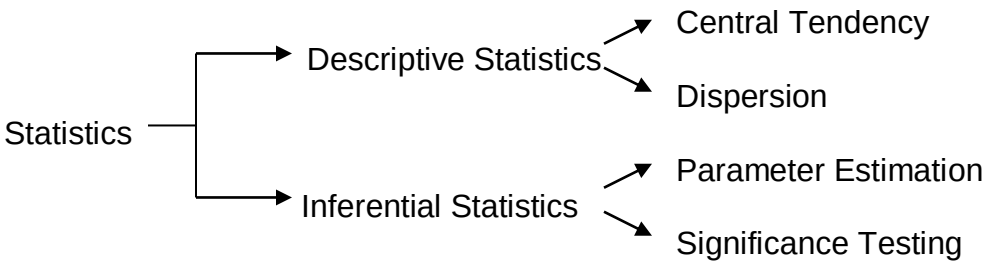
An "estimator" or "point estimate" is a statistic (that is, a function of the data) that is used to infer the value of an unknown parameter in a statistical model. The parameter being estimated is sometimes called the estimand. It can be either finite-dimensional (in parametric and semi-parametric models), or infinite-dimensional (semi-parametric and non-parametric models).

Statistical Range

In statistics, the range of a set of data is the difference between the largest and smallest values. It can give you a rough idea of how the outcome of the data set will be before you look at it actually Difference here is specific, the range of a set of data is the result of subtracting the smallest value from largest value.

However, in descriptive statistics, this concept of range has a more complex meaning. The range is the size of the smallest interval (statistics) which contains all the data and provides an indication of statistical dispersion. It is measured in the same units as the data. Since it only depends on two of the observations, it is most useful in representing the dispersion of small data sets.

Types of Statistics:



Two types of statistical methods are used in analyzing data: descriptive statistics and inferential statistics. Descriptive statistics are used to synopsize data from a sample exercising the mean or standard deviation. Inferential statistics are used when data is viewed as a subclass of a specific population.

Descriptive Statistics

Descriptive statistics refer to the analysis of the data that will help you describe, summarize, or show the data in a way that some patterns might emerge. However, you need to be aware that you shouldn't withdraw conclusions besides the data analyzed. You should be simply describing the data you got.

Despite this might not seem important, it really has a crucial part in the process since it allows you to visualize huge data in a simple and effective way.

Imagine that you wanted to analyze the performance on a test of 100 students. You might be interested in seeing the overall performance or you might be interested in looking at the spread or distribution of their marks. When you use the descriptive statistics, you should present your data by starting with a table that summarizes the group data, followed by charts and graphs. Finally, at the end, you should add the statistical commentary like the discussion of the results.

Inferential Statistics

There are many occasions when you want to analyze a specific group but you simply can't have a sample of the entire population. Unlike on the previous example, you wanted to analyze the performance of 100 students,

in this case, you might want to measure the performance of all the students in a country. Since it's not doable to collect all the data, you need to choose a smaller sample of students, which will represent all the students in that country.

And this is where the inferential statistics have their crucial role. They refer to the techniques that you use that allow you to use the samples to make generalized comments regarding the entire population. So, as you understand, it's very important to be careful when selecting the sample that represents the population. It needs to be as accurate as it can or the results won't represent the truth.
The descriptive and inferential statistics have one thing in common: they both rely on the same data. However, while the descriptive statistics only relies on this particular data, the inferential statistics relies on this data to make general conclusions about a larger population.

A descriptive statistic (in the count noun sense) is a summary statistic that quantitatively describes or summarizes features from a collection of information, while descriptive statistics (in the mass noun sense) is the process of using and analyzing those statistics. Descriptive statistics is distinguished from inferential statistics (or inductive statistics) by its aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory, and are frequently non-parametric statistics. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented. For example, in papers reporting on human subjects, typically a table is included giving the overall sample size, sample sizes in important subgroups (e.g., for each treatment or exposure group), and demographic or clinical characteristics such as the average age, the proportion of subjects of each sex, the proportion of subjects with related co-morbidities, etc.

So, which one should you choose to use? You may need to use both types of statistics and the answer depends on the purpose of your research. For example, when a company is trying to show if a new medicine will be able to help patients in the future, it's in their best interest that they use inferential statistics. If they decide to use descriptive statistics, they won't be able to withdraw any conclusions regarding the population in general but simply regarding the patients that participated in the study.

What Is Data?

Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments.

Data are characteristics or information. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects, while a datum (singular of data) is a single value of a single variable.

Data are measured, collected and reported, and analyzed, whereupon it can be visualized using graphs, images or other analysis tools. Data as a general concept refers to the fact that some existing information or knowledge is represented or coded in some form suitable for better usage or processing.

The idea of a research study is to generate data that should answer the research question being posed. Sometimes a lot of data is generated. It is important that the data collected is stored and used efficiently and it is unethical not to do this.

In academic treatments of the subject, however, data are simply units of information. Data are employed in scientific research, businesses management (e.g., sales data, revenue, profits, stock price), finance, governance (e.g., crime rates, unemployment rates, literacy rates), and in virtually every other form of human organizational activity (e.g., censuses of the number of homeless people by non-profit organizations).

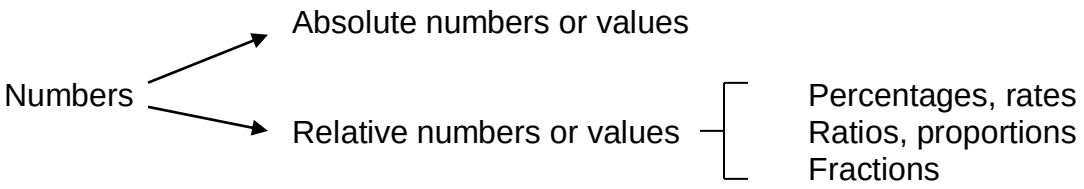
Raw data ("unprocessed data") is a collection of numbers or characters before it has been "cleaned" and corrected by researchers. Raw data needs to be corrected to remove outliers or obvious instrument or data entry errors (e.g., a thermometer reading from an outdoor Arctic location recording a tropical temperature).

Data processing commonly occurs by stages, and the "processed data" from one stage may be considered the "raw data" of the next stage. Field data is raw data that is collected in an uncontrolled "in situ" environment. Experimental data is data that is generated within the context of a scientific investigation by observation and recording.

Both descriptive and inferential statistics need to rely on some functions of the data. In the case of the descriptive statistics, it tends to rely on some classic statistics like the mean, standard deviation, min, max, skew, median, and kurtosis. In the case of the inferential statistics, they tend to use some classic statistics like the z score, t score, F-ratio, among others.

Numbers

In statistics, numbers divided into two categories:



Absolute numbers or values are the real/precise numbers, for example, 5 apples.

Relative numbers or values are dependent on other numbers. In other words, they are relative to other (absolute) numbers. Most often, those other absolute numbers are not even given. For example 2 in 5 cars drive too fast on a road. You still do not know the precise number of cars that drove too fast.

Percentages and fractions are relative.

With percentages and fractions, you do not know the precise number, only which part. So in the example above you could have used 40% or 2/5 of the cars.

Fraction looks absolute. Let's look at a carton of milk with a contents of a 1/2 litre. Is a half litre absolute? You can measure it in a measuring cylinder. Remember that when you buy a 2 litre carton of milk, the 2 is also absolute. The fraction seems to be absolute because it is a precise amount. But the fraction relates to the 'litre' behind it and is therefore relative.

Frequency

The frequency is the number of times a particular value for a variable (data item) has been observed to occur. The frequency of a value can be expressed in different ways, depending on the purpose required.

The absolute frequency describes the number of times a particular value for a variable (data item) has been observed to occur. The simplest way to express a frequency is in absolute terms.

A relative frequency describes the number of times a particular value for a variable (data item) has been observed to occur in relation to the total number of values for that variable.

The relative frequency is calculated by dividing the absolute frequency by the total number of values for the variable. Ratios, rates, proportions and percentages are different ways of expressing relative frequencies.

Percentage

A percentage expresses a value for a variable in relation to a whole population as a fraction of one hundred. The percentage total of an entire dataset should always add up to 100, as 100% represents the total, it is equal to the “whole”.

A percentage is calculated by dividing the number of times a particular value for a variable has been observed, by the total number of observations in the population, then multiplying this number by 100.

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the percentage of heads is 60% (12 divided by 20, multiplied by 100). Alternatively, the percentage of tails is 40% (8 divided by 20, multiplied by 100).

Ratio

A ratio compares the frequency of one value for a variable with another value for the variable. The first value identified in a ratio must be to the left of the colon (:) and the second value must be to the right of the colon (1st value: 2nd value).

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the ratio of heads to tails is 12:8. Alternatively, the ratio of tails to heads is 8:12.

Rate

A rate is a measurement of one value for a variable in relation to another measured quantity.

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the rate is 12 heads per 20 coin tosses. Alternatively, the rate is 8 tails per 20 coin tosses.

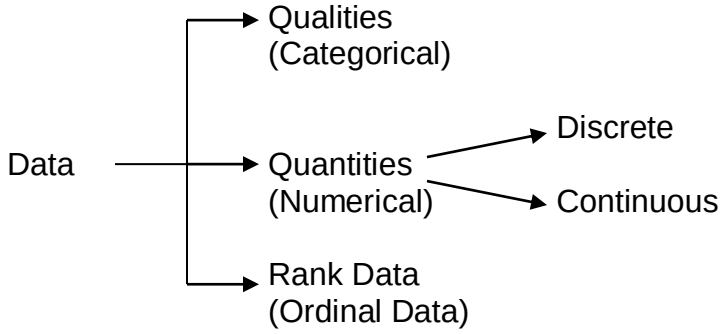
Proportion

A proportion describes the share of one value for a variable in relation to a whole. It is calculated by dividing the number of times a particular value for a variable has been observed, by the total number of values in the population.

For example, in a total of 20 coin tosses where there are 12 heads and 8 tails, the proportion of heads is 0.6 (12 divided by 20). Alternatively, the proportion of tails is 0.4 (8 divided by 20).

Data Classification

In statistics, most data fall into one of three groups: numerical, categorical, or rank data.



Qualities, Categorical Data or Enumeration Data

Qualitative properties are properties that are observed and can generally not be measured with a numerical result. They are contrasted to quantitative properties which have numerical characteristics.

Qualitative data are also a categorical data when working with statistics. Qualitative data represent characteristics such as a person's gender, marital status, hometown, or the types of movies they like. Qualitative data can take on numerical values (such as “1” indicating male and “2” indicating female), but those numbers don’t have mathematical meaning. You couldn’t add them together, therefore it could be called the enumeration data in statistics.

Categorical data is the statistical data type consisting of categorical variables or of data that has been converted into that form, for example as grouped data. More specifically, categorical data may derive from observations made of qualitative data that are summarized as counts or cross tabulations, or from observations of quantitative data grouped within given intervals. Often, purely categorical data are summarized in the form of a contingency table. However, particularly when considering data analysis, it is common to use the term "categorical data" to apply to data sets that, while containing some categorical variables, may also contain non-categorical variables.

In statistics, a categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property.

In computer science and some branches of mathematics, categorical variables are referred to as enumerations or enumerated types. Commonly (though not in this article), each of the possible values of a categorical variable is referred to as a level. The probability distribution associated with a random categorical variable is called a categorical distribution.

Examples of categorical variables:
The blood type of a person: A, B, AB or O.
The type of a rock: igneous, sedimentary or metamorphic.
The identity of a particular word (e.g., in a language model): One of V possible choices, for a vocabulary of size V.

Quantity Data or Measurement Data

Quantity, statisticians also call quantity data numerical data, is a property that can exist as a multitude or magnitude, which illustrate discontinuity and continuity. Along with analyzing its nature and classification, the issues of quantity involve such closely related topics as dimensionality, equality, proportion, the measurements of quantities, the units of measurements, number and numbering systems, the types of numbers and their relations to each other as numerical ratios. Quantity data may be also called the measurement data.

Quantitative properties have numerical characteristics. Quantities can be compared in terms of "more", "less", or "equal", or by assigning a numerical value in terms of a unit of measurement. Mass, time, distance, heat, and angular separation are among the familiar examples of quantitative properties. For example, these data have meaning as a measurement, such as a person's height, weight, IQ, or blood pressure.

In statistics, quantities can further broken into two types: Discrete Variable and Continuous Variable.

Discrete Variable Data can only take certain values. For example: the number of students in a class (you can't have half a student).

Continuous Variable Data can take any value (within a range). Example: People's heights could be any value (within the range of human heights), not just certain fixed heights. Continuous Variables would (literally) take forever to count. In fact, you would get to “forever” and never finish counting them. For example, take age. You can’t count “age”. Why not? Because it would literally take forever. For example, you could be: 25 years, 10 months, 2 days, 5 hours, 4 seconds, 4 milliseconds, 8 nanoseconds, 99 picosends and so on.

Ranked Data

Ranked data, or ordinal data, mixes numerical and categorical data. For example, rating a product on a scale from 0 (lowest) to 5 (highest) stars gives rank data. Rank data are often treated as categorical, where the groups

are ordered when graphs and charts are made. However, unlike categorical data, the numbers do have mathematical meaning. For example, if you survey 100 people and ask them to rate a product on a scale from 0 to 5, taking the average of the 100 responses will have meaning and same rank could be added up. This would not be the case with categorical data.

In statistics, ″ranking″ refers to the data transformation in which numerical or ordinal values are replaced by their rank when the data are sorted. If, for example, the numerical data 3.4, 5.1, 2.6, 7.3 are observed, the ranks of these data items would be 2, 3, 1 and 4 respectively.

Ranking the data involves putting the values in numerical order and then assigning new values to denote where in the ordered set they fall. We give the smallest value the number 1, the next largest value the number 2, the next largest number 3 etc.

The numbers 1, 2, 3... 14 that are assigned to the various values are called the ranks. If there are n values in the sample, the largest value will have rank 'n'.

Sometimes there are ties in the data. This means that two or more values are the same, so that there is no strictly increasing order. When this happens, we average the ranks for the tied values.

For example:
To rank the following sample of 14 values:

2, 34, -4, -6, 25, 2, 34, 34, 67, 28, -2, 0, 7, 23

1st Sorting the values into the order of magnitude gives:

-6, -4, –2, 0, 2, 2, 7, 23, 25, 28, 34, 34, 34, 67

2nd assigning the rank numbers:

There are 14 numbers, so the largest number has rank 14.

3rd sorting the same value numbers. The ranks 5 and 6 need to be assigned to the two '2's; hence assign rank (5+6)/2 = 5.5 to each value 2. The ranks 11,

12, and 13, need to be assigned to the three '34's, hence assign rank (11+12+13)/3 = 12 to each value 34.

Finally getting the ranks. The ranks for the sample are:

Values	-6	-4	-2	0	2	2	7	23	25	28	34	34	34	67
Ranks	1	2	3	4	5.5	5.5	7	8	9	10	12	12	12	14

Content or Values of Variables in Data

In statistics, groups of individual data points may be classified as belonging to any of various statistical data types, e.g. categorical ("red", "blue", "green"), real number (1.68, -5, 1.7e+6), odd number(1,3,5) etc. The data type is a fundamental component of the semantic content of the variable, and controls which sorts of probability distributions can logically be used to describe the variable, the permissible operations on the variable, the type of regression analysis used to predict the variable, etc. The concept of data type is similar to the concept of level of measurement, but more specific: For example, count data require a different distribution (e.g. a Poisson distribution or binomial distribution) than non-negative real-valued data require, but both fall under the same level of measurement (a ratio scale).

A categorical variable that can take on exactly two values is termed a binary variable or a dichotomous variable; an important special case is the Bernoulli variable, named after Swiss mathematician Jacob Bernoulli, the discrete probability distribution of a random variable which takes the value 1 with probability and the value 0 with probability.

Categorical variables with more than two possible values are called polytomous variables; categorical variables are often assumed to be polytomous unless otherwise specified. Discretization is treating continuous data as if it were categorical. Dichotomization is treating continuous data or polytomous variables as if they were binary variables. Regression analysis often treats category membership with one or more quantitative dummy variables.

A polychotomous variable is a variable that can have more than two values (a variable with exactly two values is called a binary variable). Polychotomous variables can be ordered, unordered, or sequential:

- Ordered polychotomous variables: variables that have some kind of order, like: "1" if you earn up to $25,000, ŏ2ō if you earn $25,001-$50,000 and ŏ3ō if you earn over $50,000.
- Unordered polychotomous variables: variables that don't have an implied order, like: "1" for male, "2" for female "3" for trans gendered male and "4" for trans gendered female.
- Sequential polychotomous variables: variables with a sequence. For example: "1" for freshmen, "2" for sophomore, "3" for junior and "4" for senior.

Polychotomous variables are usually qualitative variables, but they can be quantitative variables as well. For example, if studying birth weight of children, you could have the categories of heavy smoker/smoker/light smoker or non-smoker. But it may be more useful to code the "number of cigarettes smoked per day during pregnancy" into categories:

1 - 0 cigarettes per day.
2 - up to 5 cigarettes per day.
3 - Between 6 and 20 cigarettes per day.
4 - Over 20 cigarettes per day.

In mathematics

Magnitude (how much) and multitude (how many), the two principal types of quantities, are further divided as mathematical and physical. In formal terms, quantities their ratios, proportions, order and formal relationships of equality and inequality are studied by mathematics. The essential part of mathematical quantities consists of having a collection of variables, each assuming a set of values. These can be a set of a single quantity, referred to as a scalar when represented by real numbers, or have multiple quantities as do vectors and tensors, two kinds of geometric objects.

The mathematical usage of a quantity can then be varied and so is situationally dependent. Quantities can be used as being infinitesimal, arguments of a function, variables in an expression (independent or dependent), or probabilistic as in random and stochastic quantities. In mathematics, magnitudes and multitudes are also not only two distinct kinds of quantity but furthermore relatable to each other.

Number theory covers the topics of the discrete quantities as numbers: number systems with their kinds and relations. Geometry studies the issues of spatial magnitudes: straight lines, curved lines, surfaces and solids, all with their respective measurements and relationships.

In physical science

Establishing quantitative structure and relationships between different quantities is the cornerstone of modern physical sciences. Physics is fundamentally a quantitative science. Its progress is chiefly achieved due to rendering the abstract qualities of material entities into physical quantities, by postulating that all material bodies marked by quantitative properties or physical dimensions are subject to some measurements and observations. Setting the units of measurement, physics covers such fundamental quantities as space (length, breadth, and depth) and time, mass and force, temperature, energy, and quanta.

A distinction has also been made between intensive quantity and extensive quantity as two types of quantitative property, state or relation. The magnitude of an intensive quantity does not depend on the size, or extent, of the object or system of which the quantity is a property, whereas magnitudes of an extensive quantity are additive for parts of an entity or subsystems. Thus, magnitude does depend on the extent of the entity or system in the case of extensive quantity. Examples of intensive quantities are density and pressure, while examples of extensive quantities are energy, volume, and mass.

Processing Data

An introduction to the different types of data is the first step; it is important to establish the type of data collected in order to identify the correct means of summarizing, displaying, and analyzing that data.

Summarizing the data as frequency distributions, tables and graphs will help to identify trends (which may or may not have been expected) that may inform the subsequent analyses. Outliers may be highlighted and/or distributional tendencies (e.g., skewness) which could invalidate or overly influence results may be identified at this stage.

Ways of describing or summarizing the data are called descriptive statistics. What these aim to do is to give the relevant and useful information without losing any features of importance. Perhaps in an ideal world, all potential users of the information would have time and be capable of taking the raw data and making their own independent conclusions, thereby avoiding being at the mercy of someone else's choice of analysis. In practice, there is usually only the time or space to give/consume limited information and hence it is important that the descriptions given are accurate and we fully understand what they are and their strengths and weaknesses.

The appropriate summaries to use depend on whether the data is categorical or numeric. For numeric data there are several different potential summaries and the most appropriate depends on the distribution of the data.

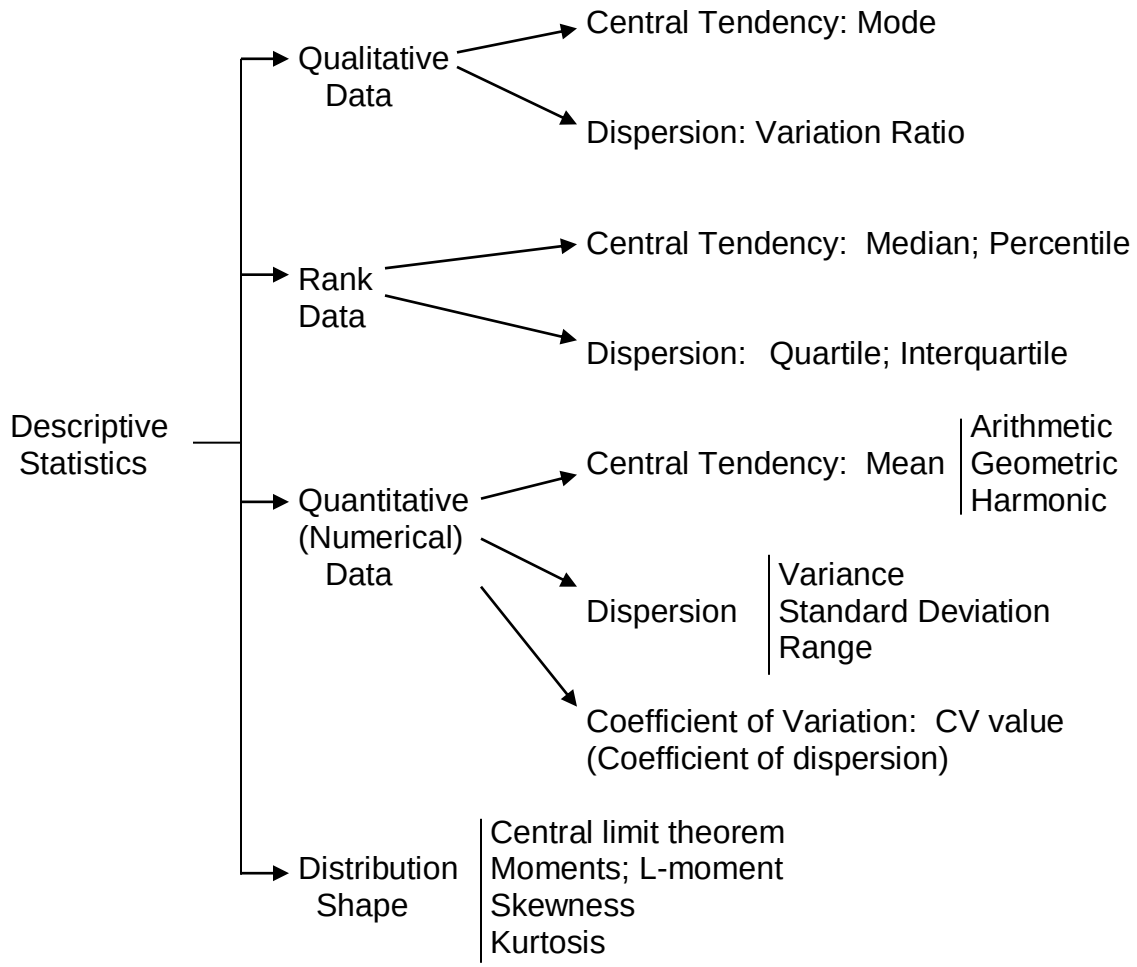
Chapter 2

Descriptive Statistics

Summary of Descriptive Statistics

Descriptive Statistics is based on the data that have initially been sorted or classified; therefore the data have been further divided as three groups, Qualitative data, Rank data and Numerical data.

Classification of Descriptive Statistics:



Central Tendency and Statistical Dispersion

In statistics, a central tendency (or measure of central tendency) is a central or typical value for a probability distribution. It may also be called a center or location of the distribution. Colloquially, measures of central tendency are often called averages.

Dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range.

The central tendency of a distribution is typically contrasted with its dispersion or variability; dispersion and central tendency are the often characterized properties of distributions. Analysis may judge whether data has a strong or a weak central tendency based on its dispersion.

Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.

The most common measures of central tendency are the arithmetic mean, the median, and the mode. A middle tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution.

Measures of Central Tendency for Qualitative Data

Mode: The mode of a set of data values is the value that appears most often. If X is a discrete random variable, the mode is the value x at which the probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

Measures of Dispersion for Qualitative Data

The variation ratio is a simple measure of statistical dispersion in nominal distributions; it is the simplest measure of qualitative variation. It is defined

as the proportion of cases which are not in the mode category: where f is the frequency of the mode, and N is the total number of cases.

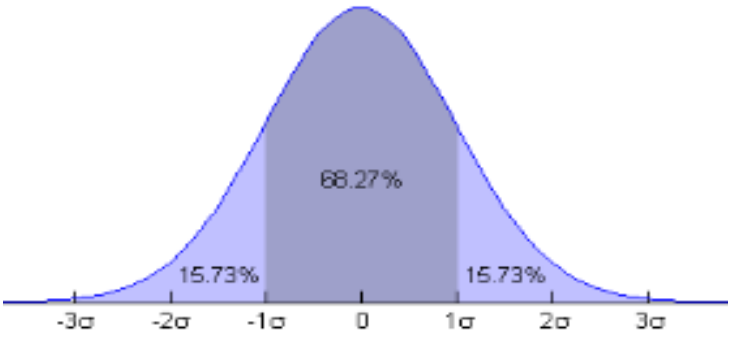
Measures of Central Tendency for Rank Data

In statistics and probability theory, a median is a value separating the higher half from the lower half of a data sample, a population or a probability distribution. For a data set, it may be thought of as "the middle" value.

Measures of Dispersion for Rank Data

Percentile

A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations falls. For example, the 20th percentile is the value below which 20% of the observations may be found. Equivalently, 80% of the observations are found above the 20th percentile.



Quartile

A quartile is a type of quantile which divides the number of data points into four more or less equal parts, or quarters. The first quartile is defined as the middle number between the smallest number and the median of the data set.

In descriptive statistics, the interquartile range (IQR), also called the midspread, middle 50%, or H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between

upper and lower quartiles, IQR = Q3 - Q1. In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data. It is a trimmed estimator, defined as the 25% trimmed range, and is a commonly used robust measure of scale.

The IQR is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that separate parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

Measures of Central Tendency for Quantitative Data

Mean

There are several kinds of mean in mathematics, especially in statistics. In mathematics and statistics, the arithmetic mean, or simply the mean or the average (when the context is clear), is the sum of a collection of numbers divided by the count of numbers in the collection.

Arithmetic mean

In mathematics and statistics, the arithmetic mean, or simply the mean or the average, is the sum of a collection of numbers divided by the count of numbers in the collection. The collection is often a set of results of an experiment or an observational study, or frequently a set of results from a survey. The term "arithmetic mean" is preferred in some contexts in mathematics and statistics, because it helps distinguish it from other means, such as the geometric mean and the harmonic mean.

The arithmetic mean may be contrasted with the median. The median is defined such that no more than half the values are larger than, and no more than half are smaller than, the median. If elements in the data increase arithmetically, when placed in some order, then the median and arithmetic average are equal. For example, consider the data sample 1,2,3,4. The average is 2.5, as is the median. However, when we consider a sample that cannot be arranged so as to increase arithmetically, such as 1,2,4,8,16, the median and arithmetic average can differ significantly. In this case, the arithmetic average is 6.2, while the median is 4. In general, the average value can vary significantly from most values in the sample, and can be larger or smaller than most of them.

For example, consider the monthly salary of 10 employees of a firm: 2500, 2700, 2400, 2300, 2550, 2650, 2750, 2450, 2600, 2400. The arithmetic mean is

(2500+2700+2400+2300+2550+2650+2750+2450+2600+2400) /10 = 2530

Geometric mean

The geometric mean is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the *n*-th root of the product of n numbers.

In mathematics, the geometric mean is a mean or average, which indicates the central tendency or typical value of a set of numbers by using the product of their values (as opposed to the arithmetic mean which uses their sum). The geometric mean is defined as the nth root of the product of n numbers, i.e., for a set of numbers x1, x2, ... xn, the geometric mean is defined as

$$\sqrt[n]{X_1 X_2 \ldots X_n}$$

For instance, the geometric mean of two numbers, say 4, 1 and 1/32, is just the cube root of their product, that is,

$$\sqrt[3]{4 \times 1 \times 1/32} = 1/2$$

Harmonic mean

The harmonic mean (sometimes called the subcontrary mean) is one of several kinds of average, and in particular, one of the Pythagorean means. Typically, it is appropriate for situations when the average of rates is desired.

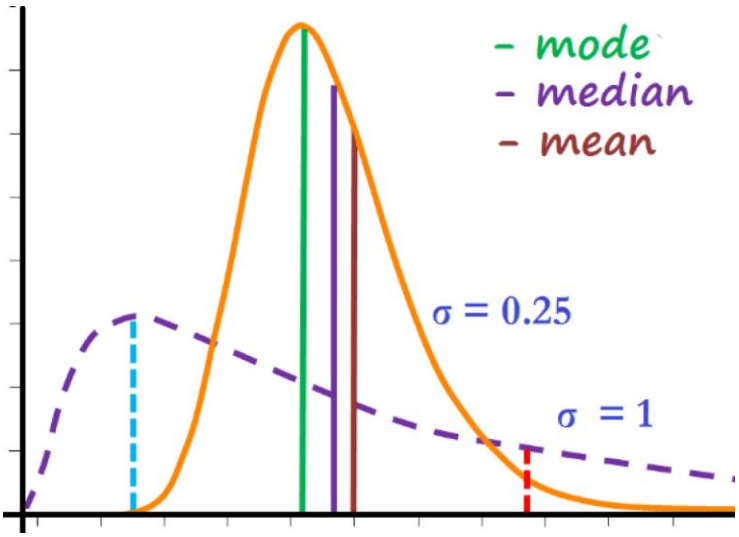
The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations. As a simple example, the harmonic mean of 1, 4, and 4 is

$$\left(\frac{1^{-1}+4^{-1}+4^{-1}}{3}\right)^{-1} = \frac{3}{\frac{1}{1}+\frac{1}{4}+\frac{1}{4}} = \frac{3}{1.5} = 2\,.$$

Difference of mode, median and mean

The arithmetic mean may be contrasted with the median. The median is defined such that no more than half the values are larger than, and no more than half are smaller than, the median. If elements in the data increase arithmetically, when placed in some order, then the median and arithmetic average are equal. For example, consider the data sample 1,2,3,4. The average is 2.5, as is the median. However, when we consider a sample that cannot be arranged so as to increase arithmetically, such as 1,2,4,8,16, the median and arithmetic average can differ significantly. In this case, the arithmetic average is 6.2, while the median is 4. In general, the average value can vary significantly from most values in the sample, and can be larger or smaller than most of them.

For instance, comparison of two log-normal distributions with equal median, but different skewness, resulting in different means and modes illustrated as follows:



Measures of Dispersion for Quantitative Data

Common examples of measures of statistical dispersion are the variance, standard deviation, and interquartile range.

Variance

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of numbers is spread out from their average value.

Standard Deviation

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Coefficient of Variation (CV)

In probability theory and statistics, the coefficient of variation (CV), also known as relative standard deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution.

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ.

$CV = \sigma / \mu$

It shows the extent of variability in relation to the mean of the population. The coefficient of variation should be computed only for data measured on a ratio scale, that is, scales that have a meaningful zero and hence allow relative comparison of two measurements (i.e., division of one measurement by the other).

Shape of a Probability Distribution

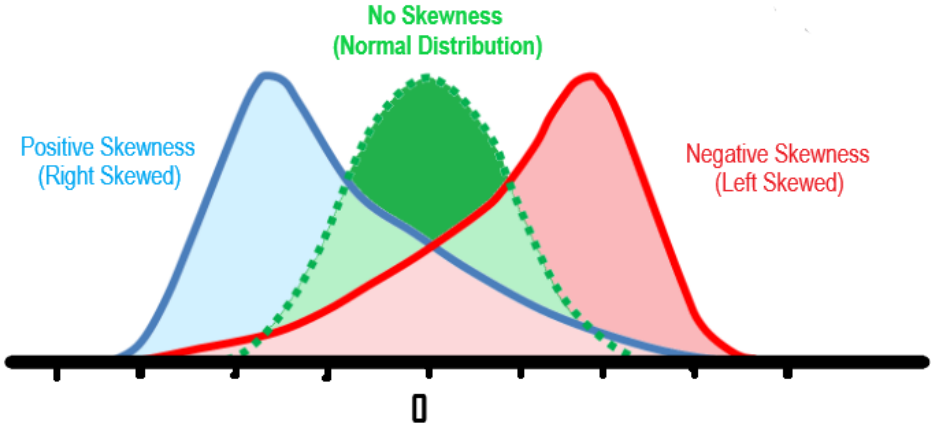
In statistics, the concept of the shape of a probability distribution arises in questions of finding an appropriate distribution to use to model the statistical properties of a population, given a sample from that population. The shape of a distribution may be considered either descriptively, using terms such as "J-shaped", or numerically, using quantitative measures such as skewness and kurtosis.

Central Limit Theorem

In probability theory, the central limit theorem (CLT) establishes that, in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

Skewness

In probability theory and statistics, skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.



Kurtosis

In probability theory and statistics, kurtosis (from Greek, kyrtos or kurtos, meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable. Like skewness, kurtosis describes the shape of a probability distribution and there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population. Different measures of kurtosis may have different interpretations.

Moment

The moments of a function are quantitative measures related to the shape of the function's graph. The concept is used in both mechanics and statistics.

L-moment

In statistics, L-moments are a sequence of statistics used to summarize the shape of a probability distribution. They are linear combinations of order statistics (L-statistics) analogous to conventional moments, and can be used to calculate quantities analogous to standard deviation, skewness and kurtosis, termed the L-scale, L-skewness and L-kurtosis respectively (the L-mean is identical to the conventional mean). Standardised L-moments are called L-moment ratios and are analogous to standardized moments. Just as for conventional moments, a theoretical distribution has a set of population L-moments. Sample L-moments can be defined for a sample from the population, and can be used as estimators of the population L-moments.

The main benefits of L-estimators are that they are often extremely simple, and often robust statistics: assuming sorted data, they are very easy to calculate and interpret, and are often resistant to outliers. They thus are useful in robust statistics, as descriptive statistics, in statistics education, and when computation is difficult. However, they are inefficient, and in modern times robust statistics M-estimators are preferred, though these are much more difficult computationally. In many circumstances L-estimators are reasonably efficient, and thus adequate for initial estimation.

Simple L-estimators can be visually estimated from a box plot, and include interquartile range, midhinge, range, mid-range, and trimean.

Count Variables

An individual piece of count data is often termed a count variable. When such a variable is treated as a random variable, the Poisson, binomial and negative binomial distributions are commonly used to represent its distribution.

Graphical Examination

Graphical examination of count data may be aided by the use of data transformations chosen to have the property of stabilising the sample variance. In particular, the square root transformation might be used when data can be approximated by a Poisson distribution (although other transformation have modestly improved properties), while an inverse sine transformation is available when a binomial distribution is preferred.

In probability theory and statistics, the index of dispersion, dispersion index, coefficient of dispersion, relative variance, or variance-to-mean ratio (VMR), like the coefficient of variation, is a normalized measure of the dispersion of a probability distribution: it is a measure used to quantify whether a set of observed occurrences are clustered or dispersed compared to a standard statistical model.

Summary of Data

Grouped data are data formed by aggregating individual observations of a variable into groups, so that a frequency distribution of these groups serves as a convenient means of summarizing or analyzing the data. There are two major types of grouping: data binning of a single-dimensional variable, replacing individual numbers by counts in bins; and grouping multi-dimensional variables by some of the dimensions (especially by independent variables), obtaining the distribution of ungrouped dimensions (especially the dependent variables).

Data Binning

Data binning (also called Discrete binning or bucketing) is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values which fall into a given small interval, a bin, are replaced by a value representative of that interval, often the central value. It is a form of quantization.

Statistical data binning is a way to group numbers of more or less continuous values into a smaller number of "bins". For example, if you have data about a group of people, you might want to arrange their ages into a smaller number of age intervals (for example, grouping every five years together). It can also be used in multivariate statistics, binning in several dimensions at once.

Histograms are an example of data binning used in order to observe underlying distributions. They typically occur in one-dimensional space and in equal intervals for ease of visualization.

Dependent and independent variables are variables in mathematical modeling, statistical modeling and experimental sciences. Dependent variables receive this name because, in an experiment, their values are studied under the supposition or hypothesis that they depend, by some law or rule (e.g., by a mathematical function), on the values of other variables. Independent variables, in turn, are not seen as depending on any other variable in the scope of the experiment in question; thus, even if the existing dependency is invertible (e.g., by finding the inverse function when it exists), the nomenclature is kept if the inverse dependency is not the object of study in the experiment.

Frequency Distribution

In statistics, a frequency distribution is a list, table or graph that displays the frequency of various outcomes in a sample. Each entry in the table contains

the frequency or count of the occurrences of values within a particular group or interval.

Contingency Table

In statistics, a contingency table (also known as a cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering and scientific research. They provide a basic picture of the interrelation between two variables and can help find interactions between them.

Correlation and Dependence

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the price of a good and the quantity the consumers are willing to purchase, as it is depicted in the so-called demand curve.

Pearson Correlation Coefficient

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and -1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation

Rank Correlation

In statistics, a rank correlation is any of several statistics that measure an ordinal association - the relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the

assignment of the ordering labels "first", "second", "third", etc. to different observations of a particular variable. A rank correlation coefficient measures the degree of similarity between two rankings, and can be used to assess the significance of the relation between them. For example, two common nonparametric methods of significance that use rank correlation are the Mann-Whitney U test and the Wilcoxon signed-rank test.

Scatter Plot

A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.

Chapter 3

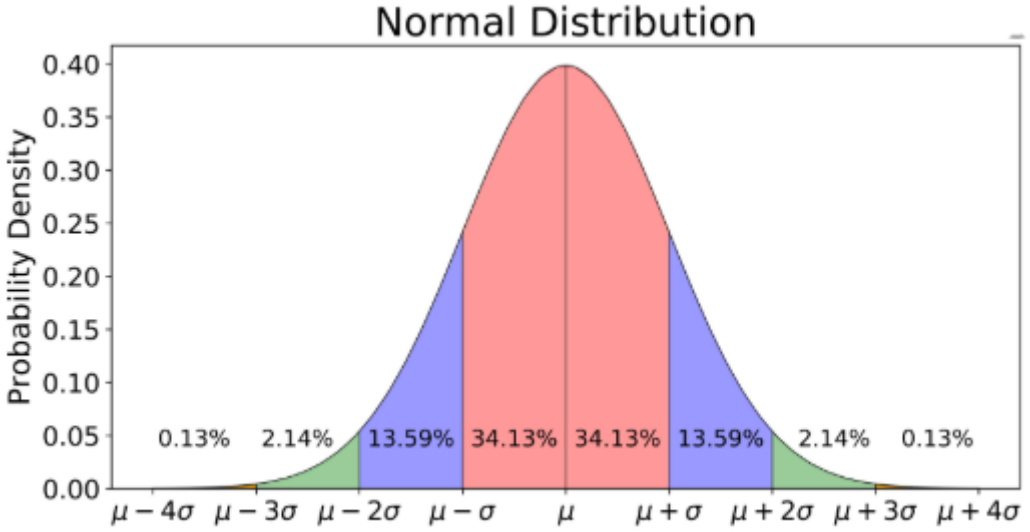
Probability Theory

Gaussian Distribution

In probability theory, a normal (or Gaussian or Gauss or Laplace-Gauss) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The parameter μ is the mean or expectation of the distribution (and also its median and mode), while the parameter σ is its standard deviation. A random variable with a Gaussian distribution is said to be normally distributed, and is called a normal deviate.

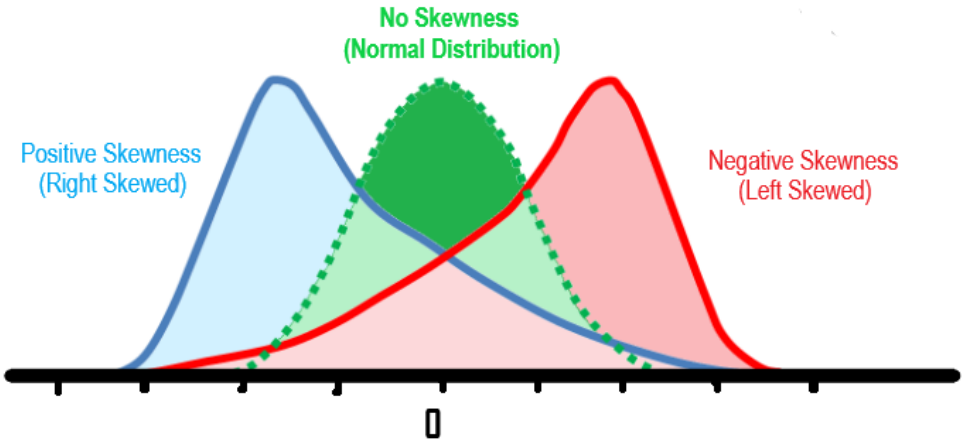


In probability theory, a normal distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is the parameter is the mean or expectation of the distribution, while the parameter is its standard deviation.

Standard Normal Distribution

The simplest case of a normal distribution is known as the standard normal distribution. This is a special case when mean μ=0 and standard deviation σ=1. The normal distribution is symmetric about its mean, and is non-zero over the entire real line. It is described by following characters:

1. The normal distribution is the only distribution whose cumulants beyond the first two (i.e., other than the mean and variance) are zero. It is also the continuous distribution with the maximum entropy for a specified mean and variance.
2. A normal distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is the mean or expectation of the distribution, while the parameter is its standard deviation.
3. Assuming that the mean and variance are finite, the normal distribution is the only distribution where the mean and variance calculated from a set of independent draws are independent of each other.
4. The normal distribution is a continuous probability distribution and a fundamental for statistics.



The graph of No Skewness green with green color showed a Normal Distribution.

The normal distribution has several implications for probability:

1. The total area under the normal curve is equal to 1.
2. The probability that a normal random variable X equals any particular value is 0.
3. The probability that X is greater than an equals the area under the normal curve bounded by a and plus infinity (as indicated by the non-shaded area in the figure below).
4. The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (as indicated by the shaded area in the figure below).

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

1. About 68% of the area under the curve falls within 1 standard deviation of the mean.
2. About 95% of the area under the curve falls within 2 standard deviations of the mean.
3. About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

To find the probability (**ρ)** associated with a normal random variable, use a graphing calculator, an online normal distribution calculator, or a normal distribution table. In the examples below, we illustrate the use of normal distribution tables. (or the probability with the z-value respectively)

ρ	z
0.8	1.28155
0.9	1.64485
0.95	1.95996
0.98	2.32635
0.99	2.57583
0.995	2.80703
0.998	3.09023

ρ	z
0.999	3.29053
0.9999	3.89059
0.99999	4.41717
0.999999	4.89164
0.9999999	5.32672
0.99999999	5.73073
0.999999999	6.10941

Calculate a range of distribution:

For example, a sample with normal distribution, and mean: $\bar{X}$=537.8, standard deviation: *S* =43.9, what is its 95% distribution?

Lowest point: $\bar{X}$ -1.96s =537.8 -1.96 x 43.9 = 451.8

Highest point: $\bar{X}$ +1.96s=537.8 +1.96 x 43.9 = 623.8

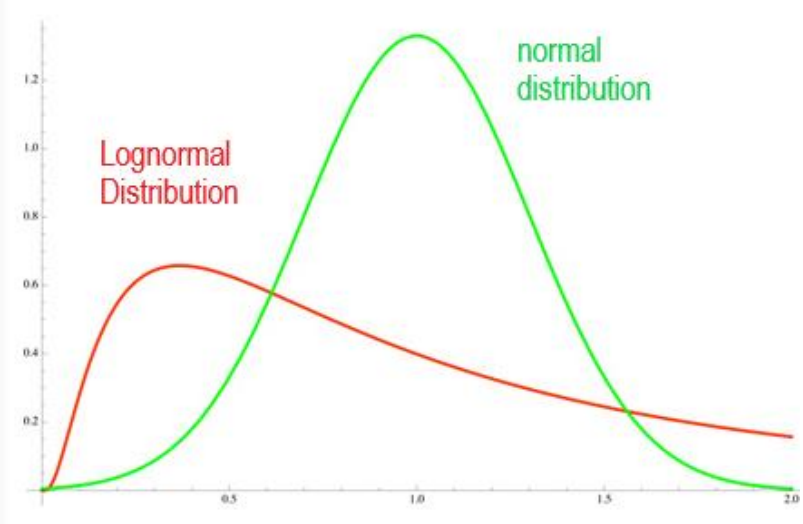
Therefore, the 95% distribution is located between 451.8 and 623.8.

The normal distribution may not be a suitable model for variables that are inherently positive or strongly skewed, such as the weight of a person or the price of a share. Such variables may be better described by other distributions, such as the log-normal distribution or the Pareto distribution.

Log-normal Distribution

In probability theory, a log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. Thus, if the random variable X is log-normally distributed, then Y = ln(X) has a normal distribution.

Equivalently, if Y has a normal distribution, then the exponential function of Y, X = exp(Y), has a log-normal distribution. A random variable which is log-normally distributed takes only positive real values. It is a convenient and useful model for measurements in exact and engineering sciences as well as medicine, economics and other fields, e.g. for energies, concentrations, lengths, financial returns and other amounts.



While most people are familiar with a normal distribution, they may not be as familiar with log-normal distribution. A normal distribution can be converted to a log-normal distribution using logarithmic mathematics. That is primarily the basis as log-normal distributions can only come from a normally distributed set of random variables.

With transformation of variance, the lognormal distribution could be in scope of application of the normal distribution. There can be a few reasons for using log-normal distributions in conjunction with normal distributions. In general most log-normal distributions are the result of taking the natural log where the base is equal to e=2.718. However, the log-normal distribution can be scaled using a different base which affects the shape of the lognormal distribution.

Overall the log-normal distribution plots the log of random variables from a normal distribution curve. In general, the log is known as the exponent to which a base number must be raised in order to produce the random variable (x) that is found along a normally distributed curve.

Pareto Distribution

The Pareto distribution, named after the Italian civil engineer, economist, and sociologist Vilfredo Pareto, is a power-law probability distribution that is used in description of social, scientific, geophysical, actuarial, and many other types of observable phenomena. Originally applied to describing the distribution of wealth in a society, fitting the trend that a large portion of wealth is held by a small fraction of the population, the Pareto distribution has colloquially become known and referred to as the Pareto principle, or "80-20 rule", and is sometimes called the "Matthew principle". This rule states that, for example, 80% of the wealth of a society is held by 20% of its population.

Quantile Function

The quantile function of a distribution is the inverse of the cumulative distribution function. The quantile function of the standard normal distribution is called the probit function, and can be expressed in terms of the inverse error function.

Sampling Error

In statistics, sampling errors are incurred when the statistical characteristics of a population are estimated from a subset, or sample, of that population. Since the sample does not include all members of the population, statistics on the sample, such as means and quartiles, generally differ from the characteristics of the entire population, which are known as parameters.

For example, if one measures the height of a thousand individuals from a country of one million, the average height of the thousand is typically not the same as the average height of all one million people in the country.

Since sampling is typically done to determine the characteristics of a whole population, the difference between the sample and population values is considered an error. Exact measurement of sampling error is generally not feasible since the true population values are unknown.

Theoretical Standard Error (SE)

The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation. If the parameter or the statistic is the mean, it is called the standard error of the mean (SEM).

The sampling distribution of a population mean is generated by repeated sampling and recording of the means obtained. This forms a distribution of different means, and this distribution has its own mean and variance. Mathematically, the variance of the sampling distribution obtained is equal to the variance of the population divided by the sample size. This is because as the sample size increases, sample means cluster more closely around the population mean.

Therefore, the relationship between the standard error of the mean and the standard deviation is such that, for a given sample size, the standard error of the mean equals the standard deviation divided by the square root of the sample size. In other words, the standard error of the mean is a measure of the dispersion of sample means around the population mean.

The standard error of the mean (SEM) can be expressed as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \text{Standard Error} = \frac{\text{Standard Deviation}}{\text{Square Root of Sample Size}}$$

where σ is the standard deviation of the population; n is the size (number of observations) of the sample.

Estimation of Standard Error of a Sample Mean

Since the population standard deviation is seldom known, the standard error of the mean is usually estimated as the sample standard deviation divided by the square root of the sample size (assuming statistical independence of the values in the sample).

$$\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$$

where ***s*** is the sample standard deviation (i.e., the sample-based estimate of the standard deviation of the population), and ***n*** is the size (number of observations) of the sample.

In those contexts where standard error of the mean is defined not as the standard deviation of the samples, but as its estimate, this is the estimate typically given as its value. Thus, it is common to see standard deviation of the mean alternatively defined as:

$$\mathrm{s}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The standard deviation of the sample mean is equivalent to the standard deviation of the error in the sample mean with respect to the true mean, since the sample mean is an unbiased estimator. Therefore, the standard error of the mean can also be understood as the standard deviation of the error in the sample mean with respect to the true mean (or an estimate of that statistic).

Note:
1. The standard error and the standard deviation of small samples tend to systematically underestimate the population standard error and standard

deviation: the standard error of the mean is a biased estimator of the population standard error.

2. In regression analysis, the term "standard error" refers either to the square root of the reduced chi-squared statistic or the standard error for a particular regression coefficient (as used in, e.g., confidence intervals).

Significance Testing

There are many different statistical significance, or hypothesis, tests. They all follow the same basic principle. The appropriate test for a given situation depends on the nature of the data being analyzed.

A typical solution:
Standardize the difference of variable and compare the standardized difference.

Null Hypothesis

All statistical significance tests start with a null hypothesis. A statistical significance test measures the strength of evidence that the data sample supplies for or against some proposition of interest.

This proposition is known as a 'null hypothesis', since it usually relates to there being 'no difference' between groups' or 'no effect' of a treatment.

The methods of null hypothesis:

For Null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{H_0}$: $\boldsymbol{\mu \geq \mu_0}$ vs. alternative hypothesis $\boldsymbol{H_1}$: $\boldsymbol{H_1}$: $\boldsymbol{\mu < \mu_0}$, it is upper/right-tailed (one tailed).

For Null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{\mu \leq \mu_0}$ vs. alternative hypothesis $\boldsymbol{H_1}$: $\boldsymbol{\mu > \mu_0}$, it is lower/left-tailed (one tailed).

For Null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$ vs. alternative hypothesis $\boldsymbol{H_1}$: $\boldsymbol{\mu \neq \mu_0}$, it is two-tailed.

For example:

CMV infected babies have the same average birthweight as non-infected babies:
It is a typical null hypothesis H_0: $\mu=\mu_0$ vs. alternative hypothesis H_1: $\mu\neq\mu_0$, it is two-tailed.

Thalassaemia does not have any effect on ferritin level:
It is a typical null hypothesis H_0: $\mu=\mu_0$ vs. alternative hypothesis H_1: $\mu\neq\mu_0$, it is two-tailed.

Even if our hypothesis is not to do with 'no difference' it is still convention that the hypothesis being tested is known as the null hypothesis.

The p-values, gives a detailed description of significance testing, and discusses the relationship between confidence intervals and significance tests.

Type I Error and Type II Error

In statistical hypothesis testing, a type I error is the rejection of a true null hypothesis (also known as a "false positive" finding or conclusion; example: "an innocent person is convicted"), while a type II error is the non-rejection of a false null hypothesis (also known as a "false negative" finding or conclusion; example: "a guilty person is not convicted"). Much of statistical theory revolves around the minimization of one or both of these errors, though the complete elimination of either is a statistical impossibility for non-deterministic algorithms. By selecting a low threshold (cut-off) value and modifying the alpha (p) level, the quality of the hypothesis test can be increased. The knowledge of Type I errors and Type II errors is widely used in medical science, biometrics and computer science.

Intuitively, type I errors can be thought of as errors of commission, and type II errors as errors of omission. For example, in the context of binary classification, when trying to decide whether an input image X is an image of a dog: an error of commission (type I) is classifying X as a dog when it isn't, whereas an error of omission (type II) is classifying X as not a dog when it is.

Statistical Background

In statistical test theory, the notion of a statistical error is an integral part of hypothesis testing. The test goes about choosing about two competing propositions called null hypothesis, denoted by H_0 and alternative hypothesis, denoted by H_1. This is conceptually similar to the judgement in a court trial. The null hypothesis corresponds to the position of defendant: just as he is presumed to be innocent until proven guilty, so is the null hypothesis presumed to be true until the data provide convincing evidence against it. The alternative hypothesis corresponds to the position against the defendant.

If the result of the test corresponds with reality, then a correct decision has been made. However, if the result of the test does not correspond with reality, then an error has occurred. There are two situations in which the decision is wrong. The null hypothesis may be true, whereas we reject H_0. On the other hand, the alternative hypothesis H_1 may be true, whereas we do not reject H_0. Two types of error are distinguished: Type I error and Type II error.

Type I Error

The first kind of error is the rejection of a true null hypothesis as the result of a test procedure. This kind of error is called a type I error and is sometimes called an error of the first kind.

In terms of the courtroom example, a type I error corresponds to convicting an innocent defendant.

Type II Error

The second kind of error is the failure to reject a false null hypothesis as the result of a test procedure. This sort of error is called a type II error and is also referred to as an error of the second kind.

In terms of the courtroom example, a type II error corresponds to acquitting a criminal.

In terms of false positives and false negatives, a positive result corresponds to rejecting the null hypothesis, while a negative result corresponds to failing to reject the null hypothesis; "false" means the conclusion drawn is incorrect. Thus, a type I error is equivalent to a false positive, and a type II error is equivalent to a false negative.

Statistical Significance

If the probability of obtaining a result as extreme as the one obtained, supposing that the null hypothesis were true, is lower than a pre-specified cut-off probability (for example, 5%), then the result is said to be statistically significant and the null hypothesis is rejected.

Chapter 4

Statistical inference

Statistical inference is the process of using data analysis to deduce properties of an underlying distribution of probability. Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.

Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (first) selecting a statistical model of the process that generates the data and (second) deducing propositions from the model.

Common Type of Statistical Inference

Inferential Statistics	Z-test (u-test) t-test Analysis of variance (ANOVA): F-test; Bartlett's test. Binominal distribution Poisson distribution Chi-square test (χ^2-test) Nonparametric Statistic Analysis (Rank sum test) Linear regression Multiple linear regression

Z-test

A Z-test (or u-test) is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Z-test tests the mean of a distribution. For each significance

level in the confidence interval, the Z-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more convenient than the Student's t-test whose critical values are defined by the sample size (through the corresponding degrees of freedom).

Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. Therefore, many statistical tests can be conveniently performed as approximate Z-tests if the sample size is large or the population variance is known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large ($n < 30$), the Student's t-test (t-test) may be more appropriate.

How to perform a Z test when T is a statistic that is approximately normally distributed under the null hypothesis is as follows:

First, estimate the expected value μ of T under the null hypothesis, and obtain an estimate s of the standard deviation of T.

Second, determine the properties of T: one tailed or two tailed.

For Null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{H_0}$: $\boldsymbol{\mu \geq \mu_0}$ vs. alternative hypothesis $\boldsymbol{H_1}$: $\boldsymbol{H_1}$: $\boldsymbol{\mu < \mu_0}$, it is upper/right-tailed (one tailed).

For Null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{\mu \leq \mu_0}$ vs. alternative hypothesis $\boldsymbol{H_1}$: $\boldsymbol{\mu > \mu_0}$, it is lower/left-tailed (one tailed).

For Null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{\mu = \mu_0}$ vs. alternative hypothesis $\boldsymbol{H_1}$: $\boldsymbol{\mu \neq \mu_0}$, it is two-tailed.

Third, calculate the standard score:

$$Z = \frac{(\bar{X} - \mu_0)}{s},$$

which one-tailed and two-tailed p-values can be calculated as $\Phi(Z)$ (for upper/right-tailed tests), Φ (-Z)(for lower/left-tailed tests) and 2Φ (-|Z|) (for two-tailed tests) where Φ is the standard normal cumulative distribution function.

Z-test example

Suppose that in a particular geographic region, the mean and standard deviation of scores on a reading test are 100 points, and 12 points, respectively. Our interest is in the scores of 55 students in a particular school who received a mean score of 96. We can ask whether this mean score is significantly lower than the regional mean-that is, are the students in this school comparable to a simple random sample of 55 students from the region as a whole, or are their scores surprisingly low?

For example, the null hypothesis:
H_0: $\mu \geq \mu_0$ (the mean score is not lower than the regional mean)
H_1: $\mu < \mu_0$ (the mean score is lower than the regional mean)

First calculate the standard error of the mean:

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{55}} = \frac{12}{7.42} = 1.62$$

Where σ is the population standard deviation.

Next calculate the z-score, which is the distance from the sample mean to the population mean in units of the standard error:

$$z = \frac{M - \mu}{\mathrm{SE}} = \frac{96 - 100}{1.62} = -2.47$$

In this example, we treat the population mean and variance as known, which would be appropriate if all students in the region were tested. When population parameters are unknown, a t test should be conducted instead.

The classroom mean score is 96, which is -2.47 standard error units from the population mean of 100. Looking up the z-score in a table of the standard normal distribution cumulative probability, we find that the probability of observing a standard normal value below -2.47 is approximately 0.5 - 0.4932 = 0.0068. This is the one-sided p-value for the null hypothesis that the 55 students are comparable to a simple random sample from the population of all test-takers. The two-sided p-value is approximately 0.014 (twice the one-sided p-value).

Another way of stating things is that with probability 1 - 0.014 = 0.986, a simple random sample of 55 students would have a mean test score within 4 units of the population mean. We could also say that with 98.6% confidence we reject the null hypothesis that the 55 test takers are comparable to a simple random sample from the population of test-takers.

The Z-test tells us that the 55 students of interest have an unusually low mean test score compared to most simple random samples of similar size from the population of test-takers.

A deficiency of this analysis is that it does not consider whether the effect size of 4 points is meaningful. If instead of a classroom, we considered a subregion containing 900 students whose mean score was 99, nearly the same z-score and p-value would be observed. This shows that if the sample size is large enough, very small differences from the null value can be highly statistically significant. See statistical hypothesis testing for further discussion of this issue.

Z-test conditions

For the Z-test to be applicable, certain conditions must be met.

Nuisance parameters should be known, or estimated with high accuracy (an example of a nuisance parameter would be the standard deviation in a one-sample location test). Z-tests focus on a single parameter, and treat all other unknown parameters as being fixed at their true values. In practice, due to Slutsky's theorem, "plugging in" consistent estimates of nuisance parameters can be justified. However if the sample size is not large enough for these estimates to be reasonably accurate, the Z-test may not perform well.

The test statistic should follow a normal distribution. Generally, one appeals to the central limit theorem to justify assuming that a test statistic varies normally. There is a great deal of statistical research on the question of when a test statistic varies approximately normally. If the variation of the test statistic is strongly non-normal, a Z-test should not be used.

If estimates of nuisance parameters are plugged in as discussed above, it is important to use estimates appropriate for the way the data were sampled. In the special case of Z-tests for the one or two sample location problem, the

usual sample standard deviation is only appropriate if the data were collected as an independent sample.

In some situations, it is possible to devise a test that properly accounts for the variation in plug-in estimates of nuisance parameters. In the case of one and two sample location problems, a t-test does this.

The z-score is often used in the z-test in standardized testing - the analog of the Student's t-test for a population whose parameters are known, rather than estimated. As it is very unusual to know the entire population, the t-test is much more widely used.

T-Test

The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance.

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. It is mostly used when the data sets, like the data set recorded as the outcome from flipping a coin 100 times, would follow a normal distribution and may have unknown variances. A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

A t-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's t distribution. The t-test can be used, for example, to determine if the means of two sets of data are significantly different from each other. (To conduct a test with three or more means, one must use an analysis of variance)

Assumptions

Most test statistics have the form t = Z/s, where Z and s are functions of the data.
Z may be sensitive to the alternative hypothesis (i.e., its magnitude tends to be larger when the alternative hypothesis is true), whereas s is a scaling parameter that allows the distribution of t to be determined.

As an example, in the one-sample t-test

$$t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$$

where $\overline{X}$ is the sample mean from a sample $X_1, X_2, \ldots X_n$, of size ***n***, ***s*** is the standard error of the mean, $\widehat{\sigma}$ (or $\overline{S}$) is the estimate of the standard deviation of the population, and μ is the population mean.

T-Test Assumptions

The first assumption made regarding t-tests concerns the scale of measurement. The assumption for a t-test is that the scale of measurement applied to the data collected follows a continuous or ordinal scale, such as the scores for an IQ test.

The second assumption made is that of a simple random sample, that the data is collected from a representative, randomly selected portion of the total population.
The third assumption is the data, when plotted, results in a normal distribution, bell-shaped distribution curve.

The final assumption is the homogeneity of variance. Homogeneous, or equal, variance exists when the standard deviations of samples are approximately equal.

T-Test Calculations

Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.

The outcome of the t-test produces the t-value. This calculated t-value is then compared against a value obtained from a critical value table (called the T-Distribution Table). This comparison helps to determine the effect of chance alone on the difference, and whether the difference is outside that chance range. The t-test questions whether the difference between the groups represents a true difference in the study or if it is possibly a meaningless random difference (hull hypothesis test).

The formula (principle) for t-value calculation is as follows:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

where the $\bar{X}$ is the mean of sample; μ_0 is the mean of the population; $s / \sqrt{n}$ is the Standard Error (SE) of the mean. It is the similar with the formula in the assumption of t-test.

T-Distribution Tables

The T-Distribution Table is available in one-tail and two-tails formats. The former is used for assessing cases which have a fixed value or range with a clear direction (positive or negative). For instance, what is the probability of output value remaining below -3, or getting more than seven when rolling a pair of dice? The latter is used for range bound analysis, such as asking if the coordinates fall between -2 and +2.

The calculations can be performed with standard software programs that support the necessary statistical functions, like those found in MS Excel.

T-Values and Degrees of Freedom

The t-test produces two values as its output: t-value and degrees of freedom. The t-value is a ratio of the difference between the mean of the two sample sets and the variation that exists within the sample sets. While the numerator value (the difference between the mean of the two sample sets) is straightforward to calculate, the denominator (the variation that exists within

the sample sets) can become a bit complicated depending upon the type of data values involved. The denominator of the ratio is a measurement of the dispersion or variability. Higher values of the t-value, also called t-score, indicate that a large difference exists between the two sample sets. The smaller the t-value, the more similarity exists between the two sample sets. (A large t-score indicates that the groups are different or a small t-score indicates that the groups are similar.)

Degrees of freedom refers to the values in a study that has the freedom to vary and are essential for assessing the importance and the validity of the null hypothesis. Computation of these values usually depends upon the number of data records available in the sample set.

Explaining the T-Test

Essentially, a t-test allows us to compare the average values of the two data sets and determine if they came from the same population. In the above examples, if we were to take a sample of students from class A and another sample of students from class B, we would not expect them to have exactly the same mean and standard deviation. Similarly, samples taken from the placebo-fed control group and those taken from the drug prescribed group should have a slightly different mean and standard deviation.

Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement by assuming a null hypothesis that the two means are equal (Null hypothesis $\boldsymbol{H_0}$: $\boldsymbol{\mu=\mu_0}$ vs. alternative hypothesis $\boldsymbol{H_1}$: $\boldsymbol{\mu\neq\mu_0}$). Based on the applicable formulas, certain values are calculated and compared against the standard values, and the assumed null hypothesis is accepted or rejected accordingly.

If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and are probably not due to chance (sampler error). The t-test is just one of many tests used for this purpose.

Statisticians must additionally use tests other than the t-test to examine more variables and tests with larger sample sizes. For a large sample size, statisticians use a z-test. Other testing options include the chi-square test and the f-test.

There are three types of t-tests, and they are categorized as dependent and independent t-tests.

KEY TAKEAWAYS

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features.

The t-test is one of many tests used for the purpose of hypothesis testing in statistics.

Calculating a t-test requires three key data values. They include the difference between the mean values from each data set (called the mean difference), the standard deviation of each group, and the number of data values of each group.
There are several different types of t-test that can be performed depending on the data and type of analysis required and they are categorized as dependent and independent t-tests.

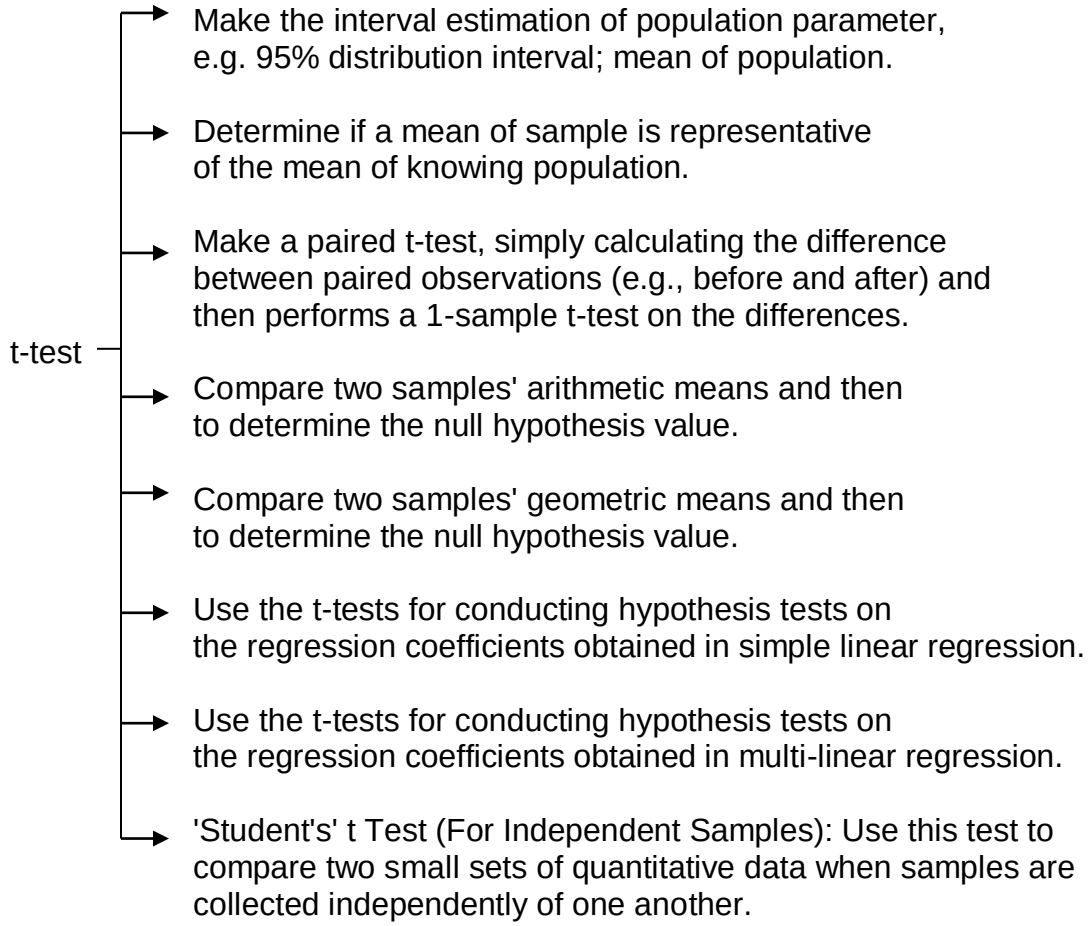
Ambiguous Test Results

Consider that a drug manufacturer wants to test a newly invented medicine. It follows the standard procedure of trying the drug on one group of patients and giving a placebo to another group, called the control group. The placebo given to the control group is a substance of no intended therapeutic value and serves as a benchmark to measure how the other group, which is given the actual drug, responds.

After the drug trial, the members of the placebo-fed control group reported an increase in average life expectancy of three years, while the members of the group who are prescribed the new drug report an increase in average life expectancy of four years. Instant observation may indicate that the drug is indeed working as the results are better for the group using the drug. However, it is also possible that the observation may be due to a chance occurrence, especially a surprising piece of luck. A t-test is useful to conclude if the results are actually correct and applicable to the entire population.

In a school, 100 students in class A scored an average of 85% with a standard deviation of 3%. Another 100 students belonging to class B scored an average of 87% with a standard deviation of 4%. While the average of class B is better than that of class A, it may not be correct to jump to the conclusion that the overall performance of students in class B is better than that of students in class A. This is because there is natural variability in the test scores in both classes, so the difference could be due to chance alone. A t-test can help to determine whether one class fared better than the other.

Applications of t-test



(The z-score is often used in the z-test in standardized testing - the analog of the Student's t-test for a population whose parameters are known, rather than estimated. As it is very unusual to know the entire population, the t-test is much more widely used.)

Example 1: Interval estimation

In statistics, interval estimation is the use of sample data to calculate an interval of possible values of an unknown population parameter; this is in contrast to point estimation, which gives a single value.

The most prevalent forms of interval estimation are: confidence intervals; and credible intervals (a Bayesian method). Other forms include: likelihood intervals; and fiducial intervals.

The t-test is used when the variable is numerical and only one population or group is being studied. It is common that the t-test is used in testing one population proportion or making the interval estimation of population parameter, e.g. 95% distribution interval; mean of population.

For example, a sample of 144 with normal distribution, and mean: $\bar{X}$=537.8, standard deviation: S=43.9, what is its mean in 95% confidence intervals?

Normally, there are n - 1 degrees of freedom (with *n* being the total number of observations). Therefore, the degrees of freedom: $\boldsymbol{v} = \boldsymbol{n} - 1 = 144 - 1 = 143$. Also, how a confidence or a probability level, e.g. α=0.05 (95%), should be considered for testing.

Use the $\boldsymbol{v}$ and a to find the t-value (t=1.979) from the t-table, and the following formula to calculate the mean of sample in 95% confidence intervals.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$= 43.9/\sqrt{144} = 3.658$

$\bar{X} - 1.979 \frac{s}{\sqrt{n}} = 537.8 - 1.979 \times 3.658 = 537.8 - 7.24 = 530.6;$

$\bar{X} + 1.979 \frac{s}{\sqrt{n}} = 537.8 + 1.979 \times 3.658 = 537.8 + 7.24 = 545.0$

Therefore, its mean in 95% confidence intervals is from 530.6 to 545.0. It's as follows:

$$(\bar{Y} - 1.979 \frac{s}{\sqrt{n}} = 530.6,\ \bar{Y} + 1.979 \frac{s}{\sqrt{n}} = 545.0) = (530.6, 545.0)$$

Example 2: Determine a mean

The t-test could be used in determining if a mean of sample is representative of the mean of knowing population. (Null hypothesis: $\mu = \mu_{(0)}$ **?**)

Suppose you have a set of sample data: $n=25$, $\bar{Y}=74.2$, $s = 6.5$; and the mean of knowing population: $\mu_0 = 72$.
Question: is the mean of the sample representative of the mean of knowing population?

Solution:

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu=72(\mu_0)$
H_1: $\mu>72(\mu_0)$
$\alpha = 0.05$ (95% confidence)

2nd Calculating the t-value:

$$t = \frac{\bar{Y} - \mu_0}{s / \sqrt{n}} = \frac{74.2 - 72}{6.5 / \sqrt{25}} = 1.692$$

3rd Find the probability (ρ) of the t-value, which is the distance from the sample mean to the population mean in units of the standard error:

With degrees of freedom, ν =n-1=25-1=24, and t-table in a table of the standard normal distribution cumulative probability, we find that the

probability (ρ) of observing a standard normal value, 1.692, is approximately: 0.10 > ρ >0.05 (two-sided p-value).

Finally, based on the confidence $\alpha = 0.05$ and $\rho > \alpha$, it indicated that the hull hypothesis tested is not refuse the $\boldsymbol{H_0}$.

We could also say that with 95% confidence we accept the null hypothesis that there are no difference between the mean of sample and the mean of population.

Example 3: Testing the Mean Difference for Paired Data

The correlated t-test is performed when the samples typically consist of matched pairs of similar units, or when there are cases of repeated measures. For example, there may be instances of the same patients being tested repeatedly - before and after receiving a particular treatment. In such cases, each patient is being used as a control sample against themselves.

This method also applies to cases where the samples are related in some manner or have matching characteristics, like a comparative analysis involving children, parents or siblings. Correlated or paired t-tests are of a dependent type, as these involve cases where the two sets of samples are related.

The formula for computing the t-value and degrees of freedom for a paired t-test is:

$$t = \frac{\bar{d} - 0}{s_{(d)} / \sqrt{n}}$$

$$S_{(d)} = \sqrt{\frac{\Sigma d^2 - (\Sigma d)^2/n}{n-1}}$$

where $\overline{d}$ is the mean of the difference between the variables of the matched pairs; *d* is the difference between the variables of the matched pairs; $S_{(d)}$ is the standard deviation of the differences of the paired data values; *n* is the sample size (the number of paired differences); *n-1* is the degrees of freedom.
Theologically, the $\mu_{(0)} = 0$, as it is supposed that the mean of population between the paired should be no difference.

Question:

Suppose we have 8 pairs of a research record as follows; is there the difference between the paired data? ($\boldsymbol{\mu_{(d)} = 0}$ **?**)

No. of Pairs (1)	Normal (2)	Treatment (3)	Difference (d) (4)=(2)-(3)	Σd^2 (5)
1	3550	2450	1100	1210000
2	2000	2400	-400	160000
3	3000	1800	1200	1440000
4	3950	3200	750	562500
5	3800	3250	550	302500
6	3750	2700	1050	1102500
7	3450	2500	950	902500
8	3050	1750	1300	1690000
Sum	26550	20050	6500	7370000

Note: the table above shown a sorted data:
n=8; $\Sigma d=6500$; $\Sigma d^2=7370000$; $\overline{d} = \Sigma d / n = 6500/8 = 812.5$.

A statistics solution:

The data are in pairs, but you're really interested only in difference for each pair. So, you take the difference between the X1 and X2 (d value in the table) for each pair, and those paired differences make up your new set of data to work with. If the paired data are the same, the average of the paired differences should be 0. If the parried data are significant differences, the t-test could provide a solution.

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu_{(d)} = 0$ (the paired data are the same)
H_1: $\mu_{(d)} \neq 0$ (the paired data are different)
$\alpha = 0.05$

2nd Calculating the t-value:

$$S_{(d)} = \sqrt{\frac{\Sigma d^2 - (\Sigma d)^2/n}{n-1}} = \sqrt{\frac{7370000-(6500)^2/8}{8-1}} = 546.25$$

$$t = \frac{\bar{d} - 0}{S_{(d)} / \sqrt{n}} = \frac{812.50 - 0}{546.25/\sqrt{8}} = 4.207$$

3rd Find the probability (ρ) of the t-value, which is the distance from the sample mean to the population mean in units of the standard error:

With degrees of freedom, ν =n-1=8-1=7, and t-table in a table of the standard normal distribution cumulative probability, we find that the probability (ρ) of observing a standard normal value, 4.207, is approximately: 0.005 > ρ >0.002 (two-sided p-value).

Finally, based on the confidence α = 0.05 and α > ρ, it indicated that the hull hypothesis tested refuse the H_0 and accept H_1.

We could also say that with 95% confidence we refuse the null hypothesis that there are differences between the mean of the paired groups.

Example 4: Compare two samples' means

This test is used when the variable is numerical and two populations or groups are being compared. Two separate random samples need to be selected, one from each population, in order to collect the data needed for this test. The null hypothesis is that the two population means are the same. The task of the test is to determine the null hypothesis: $\mu_{(1)} = \mu_{(2)}$ *or* $\mu_{(1)} - \mu_{(2)} = 0$?

A set of data as following records, X_1 and X_2, is there the difference between the two means?

No. of samples	1st sample X_1 (1)	X_1^2 (1) x (1)	2nd sample X_2 (2)	X_2^2 (2) x (2)
1	2.60	6.76	1.67	2.79
2	3.24	10.50	1.98	3.92
3	3.73	13.91	1.98	3.92
4	3.73	13.91	2.33	5.43
5	4.32	18.66	2.34	5.48
6	4.73	22.37	2.50	6.25
7	5.18	26.83	3.60	12.96
8	5.58	31.14	3.73	13.91
9	5.78	33.41	4.14	17.14
10	6.40	40.96	4.17	17.39
11	6.53	42.64	4.57	20.88
12			4.82	23.23
13			5.78	33.41
Sum	51.82	261.10	43.61	166.71

Note: the table above shown a sorted data:
n=13; $\Sigma X_1=51.82$; $\Sigma X_1^2=261.10$; $\Sigma X_2=43.61$; $\Sigma X_2^2=166.71$.

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu_{(1)} = \mu_{(2)}$
H_1: $\mu_{(1)} \neq \mu_{(2)}$
$\alpha = 0.05$

2nd The combined (or pooled) variance are involved in the calculation. The formula is combining two means and two degrees of freedom for the test:

$$S^2_c = \frac{\Sigma X_1^2 - (\Sigma X_1)^2/n_1 + \Sigma X_2^2 - (\Sigma X_2)^2/ n_2}{(n_1 - 1) + (n_2 - 1)}$$

Where S^2c is the combined sum of square, the combined sum of the square of variation, where variation is defined as the spread between each population mean.

The sum of square (S^2c) divides the combined degree of freedom then to get the combined Standard Error as follows:

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{S^2_c \left(1/n_1 + 1/ n_2\right)}$$

Therefore, the t-test statistic comparing two means is:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{[\Sigma X_1^2 - (\Sigma X_1)^2/n_1 + \Sigma X_2^2 - (\Sigma X_2)^2/n_2]}{(n_1 - 1) + (n_2 - 1)} \times (1/n_1 + 2/n_2)}}$$

3rd Sorting the variable in table and using the formula above to calculate:

$\bar{Y}_1 = \Sigma X_1/n_1 = 51.82/11 = 4.711$; $\bar{Y}_2 = \Sigma X_2/n_2 = 43.61/13 = 3.355$; and

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{4.711 - 3.355}{\sqrt{\frac{[261.10 - (51.82)^2/11 + 166.71 - (43.61)^2/13]}{(11-1) + (13-1)} \times (1/11 + 1/13)}}$$

$$= 2.539$$

Finally, find the probability (ρ) of the t-value, which is the distance from the sample mean to the population mean in units of the standard error:

With degrees of freedom, $\nu = 11+13-2=22$, and t-table in a table of the standard normal distribution cumulative probability, we find that the

probability (ρ) of observing a standard normal value, 2.539, is approximately: ρ >0.01. The ρ-value is less than α (α = 0.05). That means you do have enough evidence to reject $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Example 5: Compare two samples' geometric means

The test is similar with the test illustrated on the example four before, but it is to calculate or compare two samples' geometric means instead. The test is to determine the null hypothesis value. [$\boldsymbol{\mu_{(1)} = \mu_{(2)}}$ **?**]

In mathematics, the logarithm is the inverse function to exponentiation. That means the logarithm of a given number x is the exponent to which another fixed number, the base b, must be raised, to produce that number x. In the simplest case, the logarithm counts the number of occurrences of the same factor in repeated multiplication;

For example, since 1000 = 10x10x10 = 10^3, the "logarithm base 10" of 1000 is 3, or $\log_{10}(1000) = \log_{10}10^3 = 3$.
$\text{Log}_{10}150$ (or Lg150) is approximately 2.176, which lies between 2 and 3, just as 150 lies between $10^2 = 100$ and $10^3 = 1000$.

So, if the variable of sample is the typical geometric mean, it could use the logarithm to inverse the variable for testing.

For instance, a set of data as following records; is there the difference between the two means?

No. of samples	1st sample X_1 (1)	$\text{Lg}(X_1)$	$\Sigma(\text{Lg}X_1)^2$	2nd sample X_2 (2)	$\text{Lg}(X_2)$	$\Sigma(\text{Lg}X_2)^2$
1	40	1.60	2.5666	50	1.70	2.8865
2	20	1.30	1.6927	40	1.60	2.5666
3	30	1.48	2.1819	30	1.48	2.1819
4	25	1.40	1.9542	35	1.54	2.3841
5	10	1.00	1.0000	60	1.78	3.1618
6	15	1.18	1.3832	70	1.85	3.4044
7	25	1.40	1.9542	30	1.48	2.1819
8	30	1.48	2.1819	20	1.30	1.6927
9	40	1.60	2.5666	25	1.40	1.9542
10	10	1.00	1.0000	70	1.85	3.4044
11	15	1.18	1.3832	35	1.54	2.3841
12	30	1.48	2.1819	25	1.40	1.9542
Sum	290	16.08	22.0464	490	18.91	30.1569

Note: the table above shown a sorted data:
n=12; $\Sigma X_1=290$; $Lg(\Sigma X_1)=16.08$; $Lg(\Sigma X_1^2)=22.046$; $\Sigma X_2=490$; $Lg(\Sigma X_2)=18.91$; $Lg(\Sigma X_2^2)=30.157$.

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu_{(1)} = \mu_{(2)}$
H_1: $\mu_{(1)} \neq \mu_{(2)}$
$\alpha = 0.05$

The t-test statistic comparing two means is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{[\Sigma X_1^2-(\Sigma X_1)^2/n_1 + \Sigma X_2^2-(\Sigma X_2)^2/n_2]}{(n_1-1)+(n_2-1)} \times (1/n_1+2/n_2)}}$$

2nd To calculate it, do the following:

G1 (geometric mean of X1): $LgG1 = (\Sigma Lg X_1) / n1 = 16.0846/12=1.3404$; $Lg(X_1) = 16.08$; $\Sigma(LgX_1)^2 = 22.0464$

G2 (geometric mean of X2): $LgG2 = (\Sigma Lg X_2) / n_2 = 18.9087/12=1.5757$; $Lg(X_2) = 18.91$; $\Sigma(LgX_2)^2 = 30.1569$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{1.3404 - 1.5757}{\sqrt{\frac{[22.0454-(16.08)^2/12+30.1569-(18.91)^2/12]}{(12-1)+(12-1)} \times (1/12 + 1/12)}}$$

$$= -2.934$$

3rd Find the probability (ρ) of the t-value, which is the distance from the sample mean to the population mean in units of the standard error:

With degrees of freedom, $\nu = 12+12-2=22$, and t-table in a table of the standard normal distribution cumulative probability, we find that the

probability (ρ) of observing a standard normal value, |-2.934|, by using the absolute value to find ρ, is approximately: 0.01>ρ>0.005. The ρ-value is less than α (α = 0.05). That means you do have enough evidence to reject $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Example 6: Compare the two means of two large samples (>50 or >100)

The relationship between margin of error and sample size is simple: As the sample size increases, the margin of error decreases. Looking at formula for standard error for the sample mean:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population; n is the size (number of observations) of the sample.

You may notice that it has an n in denominator of a fraction; this is the case for most any standard error formula. As n increases, the denominator of this fraction increases, which makes the overall fraction get smaller. That makes the margin of error smaller and results in a narrower confidence interval. Therefore, you can use much simpler method to compare the means of two large samples:

$$u = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_1}{n_2}}}$$

A set of data:

1st sample: n1 =156, $\bar{X}_1$=465.13, S1=54.80

2nd sample: n2 =74, $\bar{X}_2$=422.16, S2=44.20

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu_{(1)} = \mu_{(2)}$
H_1: $\mu_{(1)} \neq \mu_{(2)}$
$\alpha = 0.05$

2nd Using the following simple formula of the combined (or pooled) variance for the calculation, the formula is combining two means and two degrees of freedom for the test:

$$u = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_1}{n_2}}} = \frac{465.13 - 422.16}{\sqrt{\frac{(54.80)^2}{156} + \frac{(44.20)^2}{74}}}$$

$= 6.360$

3rd Find the probability (ρ) of the t-value, which is the distance from the sample mean to the population mean in units of the standard error:

With degrees of freedom, $\nu = \infty$, and t-table in a table of the standard normal distribution cumulative probability, we find that the probability (ρ) of observing a standard normal value, 6.360, is approximately: $0.001 > \rho$. The ρ-value is less than α ($\alpha = 0.05$). That means you do have enough evidence to reject H_0 and accept H_1.

Homoscedasticity

In statistics, a sequence of random variables is homoscedastic if all its random variables have the same finite variance. This is also known as homogeneity of variance. The complementary notion is called heteroscedasticity. The spellings homoskedasticity and heteroskedasticity are also frequently used.

The t-test analyses require homoscedasticity, otherwise the adjusted t'-test may be applied.

Example 7: Calculation and test of homoscedasticity

Suppose we have a set of data (one sample is much more than another) as follows:

Sample 1: n1 =10, $\bar{X}_1$=6.21, S1=1.79 (S1: variance of sample 1)

Sample 2: n2 =50, $\bar{X}_2$=4.34, S2=0.56 (S2: variance of sample 2)

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu_{(1)} = \mu_{(2)}$

H_1: $\mu_{(1)} \neq \mu_{(2)}$

$\alpha = 0.05$

2nd Using the F-test to compare the two variances.

$F = S^2_1 / S^2_1 = (1.79)^2 / (0.56)^2 = 10.22$

Degree of freedom: $\nu_1 = 10-1 = 9$, $\nu_2 = 50-1 = 49$; Find the probability (ρ) of the f-value from F-table: 0.05 > ρ, and with $\alpha = 0.05$, $\alpha >$ ρ, refuse the H_0 and accept H_1. It indicated that the two samples don't have the same finite variance.

3rd Using the following formula to get t' value and calculating the adjusted $t_{0.05}$ value.

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_1}{n_2}}}$$

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_1}{n_2}}} = \frac{6.21 - 4.34}{\sqrt{\frac{(1.79)^2}{10} + \frac{(0.56)^2}{50}}}$$

$$= 3.272$$

Based on the testing level of α ($\alpha = 0.05$), from t-table with $\nu_1 = 9$, $\nu_2 = 49$, find the value: $t_{0.05,9} = 2.262$, $t_{0.05,49} = 2.009$.

To make an adjustment by multiply $t_{0.05,9} = 2.262$, $t_{0.05,49} = 2.009$ (as a weight power factor) with S^2_{X1}, S^2_{X2} respectively:

$S^2_{\bar{X}_1} = (1.79)^2/10 = 0.3204$

$S^2_{\bar{X}_2} = (0.56)^2/50 = 0.006272$

$$t'_{0.05} = \frac{(S^2_{\bar{X}_1})(t_{0.05,X_1}) + (S^2_{\bar{X}_2})(t_{0.05,x2})}{S^2_{\bar{X}_1} + S^2_{\bar{X}_2}}$$

$$t'_{0.05} = \frac{(0.3204)(2.262) + (0.006272)(2.009)}{0.3204 + 0.006272} = 2.257$$

Based on t' (3.272) $>$ $t'_{0.05}$ (2.257); then $0.05 > \rho$; ($\alpha = 0.05$, from t-table with $\nu_1 = 9$, $\nu_2 = 49$, find the value: $t_{0.05,9} = 2.262$, $t_{0.05,49} = 2.009$)

With $\alpha = 0.05$; reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$. That means you do have enough evidence to reject $\boldsymbol{H_0}$ – difference between the two sample means.

Chapter 5

Analysis of Variance (ANOVA)

An analysis of variance (ANOVA) is the synthesis of several ideas and it is used for multiple purposes. As a consequence, it is difficult to define concisely or precisely. To conduct a test with three or more means, one must use ANOVA. ANOVA is a collection of statistical models and their associated estimation procedures used to analyze the differences among group means in a sample, and an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not.

ANOVA is a form of statistical hypothesis testing heavily used in the analysis of experimental data. A test result (calculated from the null hypothesis and the sample) is called statistically significant if it is deemed unlikely to have occurred by chance, assuming the truth of the null hypothesis. A statistically significant result, when a probability (p-value) is less than a pre-specified threshold (significance level), justifies the rejection of the null hypothesis, but only if the a priori probability of the null hypothesis is not high.

In the typical application of ANOVA, the null hypothesis is that all groups are random samples from the same population. For example, when studying the effect of different treatments on similar samples of patients, the null hypothesis would be that all treatments have the same effect (perhaps none). Rejecting the null hypothesis is taken to mean that the differences in observed effects between treatment groups are unlikely to be due to random chance.

The terminology of ANOVA is largely from the statistical design of experiments. The experimenter adjusts factors and measures responses in an attempt to determine an effect. Factors are assigned to experimental units by a combination of randomization and blocking to ensure the validity of the results. Blinding keeps the weighing impartial. Responses show a variability that is partially the result of the effect and is partially random error.

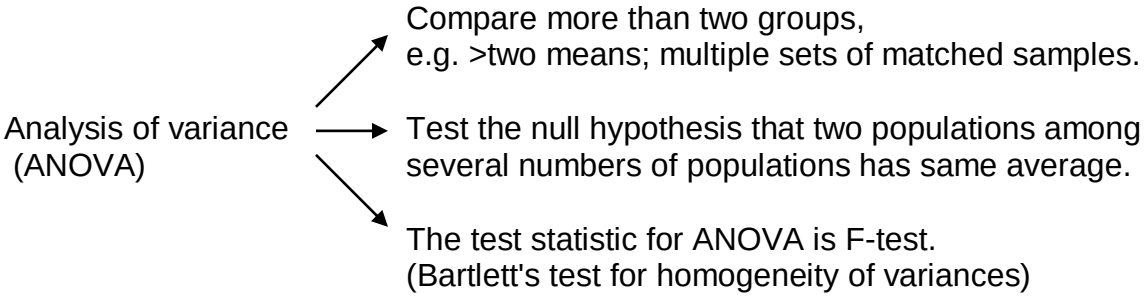
"Classical" ANOVA

"Classical" ANOVA for balanced data does three things at once:

1. As exploratory data analysis, an ANOVA employs an additive data decomposition, and its sums of squares indicate the variance of each component of the decomposition (or, equivalently, each set of terms of a linear model).
2. Comparisons of mean squares, along with an F-test ... allow testing of a nested sequence of models.
3. Closely related to the ANOVA is a linear model fit with coefficient estimates and standard errors.

In short, ANOVA is a statistical tool used in several ways to develop and confirm an explanation for the observed data. Additionally, it is computationally elegant and relatively robust against violations of its assumptions.

ANOVA provides strong (multiple sample comparison) statistical analysis. It has been adapted to the analysis of a variety of experimental designs. As a result: " ANOVA " is probably the most useful technique in the field of statistical inference."



Bartlett's test

Bartlett's test for homogeneity of variances is used to test that variances are equal for all samples. It checks that the assumption of equal variances is true before running certain statistical tests like the One-Way ANOVA. It's used when you're fairly certain your data comes from a normal distribution. Bartlett's test is used to test the null hypothesis, H0 that all k population variances are equal against the alternative that at least two are different.

Assumptions

ANOVA has been studied from several approaches, the most common of which uses a linear model that relates the response to the treatments and blocks. Note that the model is linear in parameters but may be nonlinear across factor levels. Interpretation is easy when data is balanced across factors but much deeper understanding is needed for unbalanced data.

The analysis of variance can be presented in terms of a linear model, which makes the following assumptions about the probability distribution of the responses:

Independence of observations - this is an assumption of the model that simplifies the statistical analysis.
Normality - the distributions of the residuals are normal.
Equality (or "homogeneity") of variances, called homoscedasticity - the variance of data in groups should be the same.

Summary of assumptions
The normal-model based ANOVA analysis assumes the independence, normality and homogeneity of variances of the residuals. The randomization-based analysis assumes only the homogeneity of the variances of the residuals (as a consequence of unit-treatment additivity) and uses the randomization procedure of the experiment. Both these analyses require homoscedasticity, as an assumption for the normal-model analysis and as a consequence of randomization and additivity for the randomization-based analysis.

However, studies of processes that change variances rather than means (called dispersion effects) have been successfully conducted using ANOVA. There are no necessary assumptions for ANOVA in its full generality, but the F-test used for ANOVA hypothesis testing has assumptions and practical limitations which are of continuing interest.

Problems which do not satisfy the assumptions of ANOVA can often be transformed to satisfy the assumptions. The property of unit-treatment additivity is not invariant under a "change of scale", so statisticians often use transformations to achieve unit-treatment additivity. If the response variable

is expected to follow a parametric family of probability distributions, then the statistician may specify (in the protocol for the experiment or observational study) that the responses be transformed to stabilize the variance. Also, a statistician may specify that logarithmic transforms be applied to the responses, which are believed to follow a multiplicative model. According to Cauchy's functional equation theorem, the logarithm is the only continuous transformation that transforms real multiplication to addition.

Characteristics

ANOVA is used in the analysis of comparative experiments, those in which only the difference in outcomes is of interest. The statistical significance of the experiment is determined by a ratio of two variances. This ratio is independent of several possible alterations to the experimental observations: Adding a constant to all observations does not alter significance. Multiplying all observations by a constant does not alter significance. So ANOVA statistical significance result is independent of constant bias and scaling errors as well as the units used in expressing observations. In the era of mechanical calculation it was common to subtract a constant from all observations (when equivalent to dropping leading digits) to simplify data entry. This is an example of data coding.

The calculations of ANOVA can be characterized as computing a number of means and variances, dividing two variances and comparing the ratio to a handbook value to determine statistical significance. Calculating a treatment effect is then trivial: "the effect of any treatment is estimated by taking the difference between the mean of the observations which receive the treatment and the general mean".

The fundamental technique is a partitioning of the total sum of squares SS into components related to the effects used in the model, for example, the model for a simplified ANOVA with one type of treatment at different levels.

$$SS_{\text{total}} = SS_{\text{error}} + SS_{\text{treatment}}$$

$$SS_{\text{total}} = \Sigma X^2 - (\Sigma X)^2/\mathbf{n}$$

The F-test

The F-test is used for comparing the factors of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F test statistic as follows.

$$F = \frac{\text{Variance between the treatments}}{\text{Variance within the treatments}} = \frac{\text{MS treatments}}{\text{MS error}} = \frac{\text{SS treatments} / (k - 1)}{\text{SS error} / (N - k)}$$

Where, MS is mean square, k = number of treatments and n = total number of cases to the F-distribution with k - 1, n - k degrees of freedom. Using the F-distribution is a natural candidate because the test statistic is the ratio of two scaled sums of squares each of which follows a scaled chi-squared distribution.

Example 1: An analysis of variance for more than two means

Suppose we have the three random samples, each group with six individual numbers as follows:

3 groups of treatments (samples)

	X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	ΣX $X_1+X_2+X_3$	ΣX^2 $X_1^2+X_2^2+X_3^2$
1	3.3	10.89	4.4	19.36	3.60	12.96	11.3	43.21
2	3.6	12.96	4.4	19.36	4.40	19.36	12.4	51.68
3	4.3	18.49	3.4	11.56	5.10	26.01	12.8	56.06
4	4.1	16.81	4.2	17.64	5.00	25.00	13.3	59.45
5	4.2	17.64	4.7	22.09	5.50	30.25	14.4	69.98
6	3.3	10.89	4.2	17.64	4.70	22.09	12.2	52.62
ΣX	22.80		25.30		28.3		76.40	
ΣX^2		87.68		107.65		135.67		331.00
$(\Sigma X)^2$	519.84		640.09		800.89		5836.96	
N or n	6		6		6		18	
$(\Sigma X)^2/n$	86.640		106.682		133.482		324.276	

Note: the table above showed a sorted data.

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu_{(1)} = \mu_{(2)} = \mu_{(3)}$
H_1: $\mu_{(1)} \neq \mu_{(2)} \neq \mu_{(3)}$
$\alpha = 0.05$

2nd To calculate it, do the following:

$SS_{total} = \Sigma X^2 - (\Sigma\Sigma X)^2/N$
$= (87.68+107.65+135.67) - (76.40)^2/18$
$= 331.00 - 324.276$
$= 6.724$

$SS_{treat} = \Sigma(\Sigma X)^2/n_i - (\Sigma X)^2/N$
$= (519.84+640.09+800.89)/6 - (76.40)^2/18$
$= 326.804 - 324.276$
$= 2.527$

$SS_{error} = SS_{total} - SS_{treat}$
$= 6.724 - 2.528$
$= 4.197$

$MS_{treat} = SS_{treat} / (k-1) = 2.527 / (3-1) = 1.263$

$MS_{error} = SS_{error} / (N-k) = 4.197 / (18-3) = 0.280$

$$F = \frac{MS_{treat}}{MS_{error}} = \frac{1.263}{0.280} = 4.511$$

3rd Find the probability (ρ) of the F-value:

With degrees of freedom, $\nu_i = 3 - 1 = 2$, $\nu_j = 18-3 = 15$, and F-table, we find that the probability (ρ) of observing a standard normal value, 4.511, is approximately: 0.05>ρ>0.01. The ρ-value is less than α (0.05) or $\alpha > \rho$. That means you do have enough evidence to reject H_0 and accept H_1.

It means that there are the statistic significant differences among the groups, but it doesn't tell us the difference between each other.

Example 2: An analysis of variance for the placebo data

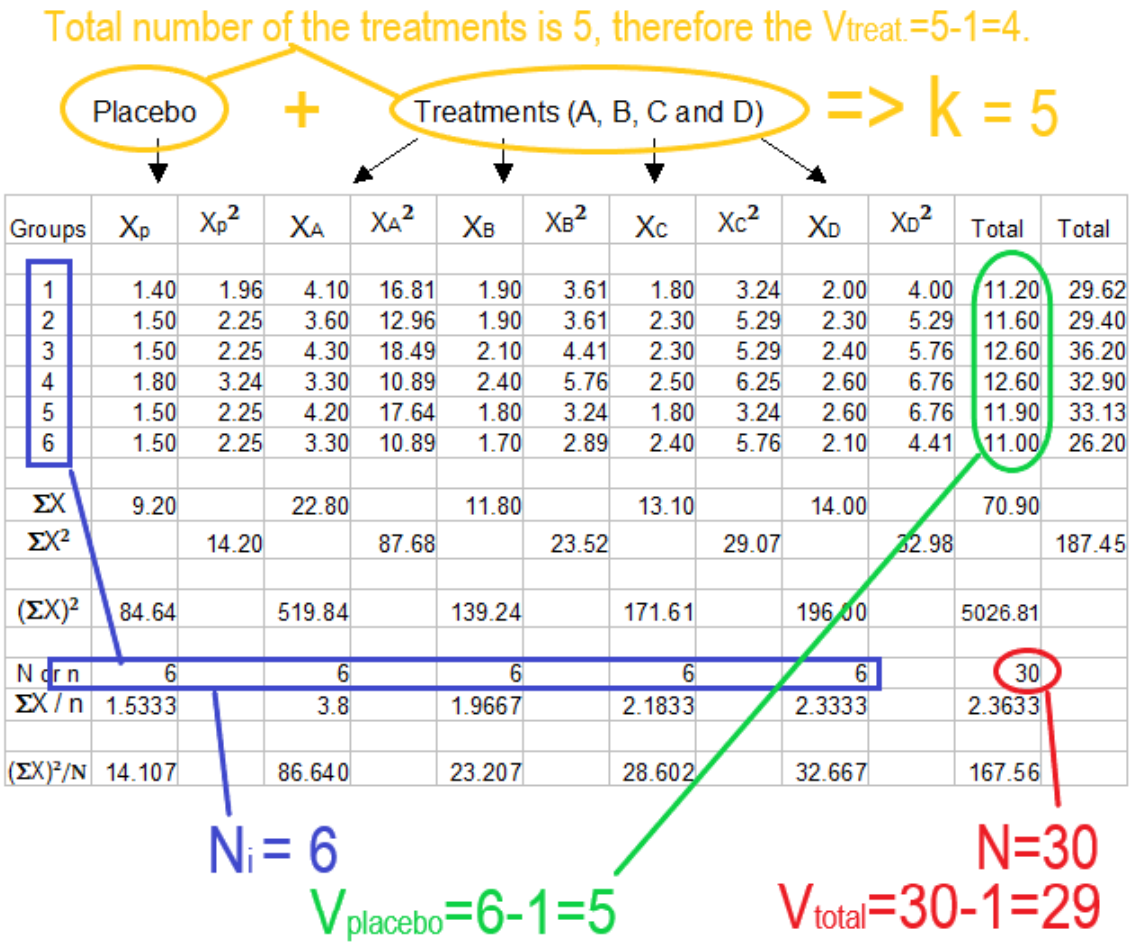
Suppose we have a set of data: three random sample, each group with six individual numbers, as follows:

There 5 individual in each group, 1 individual is a placebo, with a total 6 groups in total 30 individual.

	Placebo		Treatments (A, B, C and D)									
Groups	X_p	X_p^2	X_A	X_A^2	X_B	X_B^2	X_C	X_C^2	X_D	X_D^2	Total	Total
1	1.40	1.96	4.10	16.81	1.90	3.61	1.80	3.24	2.00	4.00	11.20	29.62
2	1.50	2.25	3.60	12.96	1.90	3.61	2.30	5.29	2.30	5.29	11.60	29.40
3	1.50	2.25	4.30	18.49	2.10	4.41	2.30	5.29	2.40	5.76	12.60	36.20
4	1.80	3.24	3.30	10.89	2.40	5.76	2.50	6.25	2.60	6.76	12.60	32.90
5	1.50	2.25	4.20	17.64	1.80	3.24	1.80	3.24	2.60	6.76	11.90	33.13
6	1.50	2.25	3.30	10.89	1.70	2.89	2.40	5.76	2.10	4.41	11.00	26.20
ΣX	9.20		22.80		11.80		13.10		14.00		70.90	
ΣX^2		14.20		87.68		23.52		29.07		32.98		187.45
$(\Sigma X)^2$	84.64		519.84		139.24		171.61		196.00		5026.81	
N or n	6		6		6		6		6		30	
ΣX / n	1.5333		3.8		1.9667		2.1833		2.3333		2.3633	
$(\Sigma X)^2/N$	14.107		86.640		23.207		28.602		32.667		167.56	

Note: the table above shown a sorted data.

For more details in regarding to data sorting, you may view the illustrations as follows:



Groups	X_p	X_p^2	X_A	X_A^2	X_B	X_B^2	X_C	X_C^2	X_D	X_D^2	Total	Total
1	1.40	1.96	4.10	16.81	1.90	3.61	1.80	3.24	2.00	4.00	11.20	29.62
2	1.50	2.25	3.60	12.96	1.90	3.61	2.30	5.29	2.30	5.29	11.60	29.40
3	1.50	2.25	4.30	18.49	2.10	4.41	2.30	5.29	2.40	5.76	12.60	36.20
4	1.80	3.24	3.30	10.89	2.40	5.76	2.50	6.25	2.60	6.76	12.60	32.90
5	1.50	2.25	4.20	17.64	1.80	3.24	1.80	3.24	2.60	6.76	11.90	33.13
6	1.50	2.25	3.30	10.89	1.70	2.89	2.40	5.76	2.10	4.41	11.00	26.20
ΣX	9.20		22.80		11.80		13.10		14.00		70.90	
ΣX^2		14.20		87.68		23.52		29.07		32.98		187.45
$(\Sigma X)^2$	84.64		519.84		139.24		171.61		196.00		5026.81	
N or n	6		6		6		6		6		30	
ΣX / n	1.5333		3.8		1.9667		2.1833		2.3333		2.3633	
$(\Sigma X)^2/N$	14.107		86.640		23.207		28.602		32.667		167.56	

1st Making the null hypothesis and setting the test confidence level:

H_0: $\mu_{(1)} = \mu_{(2)} = \mu_{(3)} = \mu_{(4)} = \mu_{(5)}$ Or ($\mu_{(i)} = 0$)
H_1: $\mu_{(1)} \neq \mu_{(2)} \neq \mu_{(3)} \neq \mu_{(4)} \neq \mu_{(5)}$ Or ($\mu_{(i)} \neq 0$)
$\alpha = 0.05$

2nd To calculate it, do the following:

SS total = $\Sigma X^2 - (\Sigma X)^2/N$
SS treat = $\Sigma(\Sigma X)^2/n_i - (\Sigma X)^2/N$
SS placebo = $\Sigma(\Sigma X)^2/k - (\Sigma X)^2/N$

SS total = $\Sigma X^2 - (\Sigma X)^2/N$

$= (14.20+87.68+23.52+29.07+32.98) - (70.90)^2 / 30$

$= 187.450 - 167.560$

$= 19.890$

$SS_{treat} = \Sigma(\Sigma X)^2/n_i - (\Sigma X)^2/N$

$= (9.20^2+22.80^2+11.80^2+13.10^2+14.00^2)/6 - (70.90)^2/30$

$= (84.64+519.84+139.24+171.61+196.00)/6 - 5026.810/30$

$= 185.222 - 167.560$

$= 17.662$

$SS_{placebo} = \Sigma(\Sigma X)^2/k - (\Sigma X)^2/N$

$= (11.20^2+11.60^2+12.60^2+12.60^2+11.90^2+11.00^2)/(6-1) - (70.90)^2/30$

$= (125.44+134.56+158.76+158.76+141.61+121.00)/5 - 167.560$

$= 168.026 - 167.560$

$= 0.466$

$SS_{error} = SS_{total} - SS_{treat} - SS_{placebo}$

$= 19.890 - 17.662 - 0.466$

$= 1.762$

$\nu_{error} = \nu_{total} - \nu_{treat} - \nu_{placebo} = (30 - 1) - (5-1) - (6 - 1) = 20$

$MS_{treat} = SS_{treat} / \nu_{treat} = 17.662 / (5 - 1) = 4.416$

$MS_{placebo} = SS_{placebo} / \nu_{placebo} = 0.466 / (6 - 1) = 0.0932$

$MS_{error} = SS_{error} / \nu_{error} = 1.762 / 20 = 0.0881$

$$F = \frac{MS_{placebo}}{MS_{error}} = \frac{0.0932}{0.0881} = 1.058$$

3rd Find the probability (ρ) of the F-value:

With degrees of freedom, $\nu_{placebo} = 5$, $\nu_{error} = 20$; and F-table, we find that the probability (ρ) of observing a standard normal value, 1.058, is approximately: ρ>0.05. $\alpha = 0.05$, and ρ>α. That means you don't have enough evidence to reject $\boldsymbol{H_0}$ and therefore accept $\boldsymbol{H_0}$.

As the MS placebo and MS error have no significant statistic meaning, it is needed to make further testing – error-pooled.

4th Calculate error-pooled testing:

$SS_{error\text{-}pooled} = SS_{error} + SS_{placebo} = 1.762 + 0.466 = 2.228$

$\nu_{error\text{-}pooled} = \nu_{error} + \nu_{placebo} = 20 + 5 = 25$

$MS_{error\text{-}pooled} = SS_{error\text{-}pooled} / \nu_{error\text{-}pooled} = 2.228 / 25 = 0.0891$

$$F = \frac{MS_{treat}}{MS_{error\text{-}pooled}} = \frac{4.416}{0.0891} = 49.562$$

With degrees of freedom, $\nu_{error\text{-}pooled} = 25$, and F-table, we find that the probability (ρ) of F-value, 49.562, is approximately: 0.01>ρ>. $\alpha = 0.05$, and α > ρ. That means you do have enough evidence to reject $\boldsymbol{H_0}$ and therefore accept $\boldsymbol{H_1}$- there are statistic significant difference among the groups.

Newman-Keuls Method (q-test)

The Newman-Keuls or Student-Newman-Keuls (SNK), simply q-test, method is a stepwise multiple comparisons procedure used to identify sample means that are significantly different from each other. This procedure is often used as a post-hoc test whenever a significant difference between three or more sample means has been revealed by an analysis of variance (ANOVA). The q-test method uses different critical values for different pairs of mean comparisons. Thus, the procedure is more likely to

reveal significant differences between group means and to commit type I errors by incorrectly rejecting a null hypothesis when it is true.

Required assumptions

The assumptions of the q-test are essentially the same as for an independent groups t-test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematic than in the two-sample case since the standard error of the mean (SEM) is based on data from all groups. The assumption of independence of observations is important and should not be violated.

Procedures

The q-test method employs a stepwise approach when comparing sample means. Prior to any mean comparison, all sample means are rank-ordered in ascending or descending order, thereby producing an ordered range (p) of sample means. A comparison is then made between the largest and smallest sample means within the largest range. Assuming that the largest range is four means (or $p = 4$), a significant difference between the largest and smallest means as revealed by the q-test method would result in a rejection of the null hypothesis for that specific range of means. The next largest comparison of two sample means would then be made within a smaller range of three means (or $p = 3$). Unless there is no significant difference between two sample means within any given range, this stepwise comparison of sample means will continue until a final comparison is made with the smallest range of just two means. If there is no significant difference between the two sample means, then all the null hypotheses within that range would be retained and no further comparisons within smaller ranges are necessary.

The q-test formula:

$$q = \frac{|\bar{X}_A - \bar{X}_B|}{S_{\bar{X}_A - \bar{X}_B}} = \frac{|\bar{X}_A - \bar{X}_B|}{\sqrt{\frac{MS_{error}}{2} \times \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

where q represents the studentized range value, $\bar{X}_A$ and $\bar{X}_B$ are the largest and smallest sample means within a range, MS_{error} (standard error of the mean, SEM) is the error variance taken from the ANOVA table, and n is the sample size (number of observations within a sample).

Example 3: Compare each of two means among the multiple samples

Based on the results from the test in Example 2 before, the further q-test method for comparing the each other and finding the statistic significant difference between the groups can be conducted.

We have calculated the following 5 means from the five samples in Example 2.

ΣX / n	1.5333	3.800	1.9667	2.1833	2.3333

1st Rank the means from higher to lower and sort the data:

ΣX / n	3.800	2.333	2.183	1.967	1.533	

The table below has showed the chance of comparing with each other:

		#1	#2	#3	#4	#5
		3.800	2.333	2.183	1.967	1.533
#1	3.800		√	√	√	√
#2	2.333			√	√	√
#3	2.183				√	√
#4	1.967					√
#5	1.533					

Or based on the following calculation:

$$C_5^2 = \frac{5!}{2!\,(5-2)!} = 10$$

Calculate the difference between the two means compared: $\bar{Y}_A - \bar{Y}_B$

		#1	#2	#3	#4	#5
		3.800	2.333	2.183	1.967	1.533
#1	3.800		1.467	1.617	1.833	2.267
#2	2.333			0.150	0.366	0.800
#3	2.183				0.216	0.650
#4	1.967					0.434
#5	1.533					

e.g. $\bar{Y}_{(\#1)} - \bar{Y}_{(\#2)} = 3.800-2.333=1.467$; $\bar{Y}_{(\#1)} - \bar{Y}_{(\#3)} = 3.800-2.183=1.617$.

2nd To calculate it, do the following:

From the 4-th step of example 2:

$SS_{\text{error-pooled}} = SS_{\text{error}} + SS_{\text{placebo}} = 1.762 + 0.466 = 2.228$

$\nu_{\text{error-pooled}} = \nu_{\text{error}} + \nu_{\text{placebo}} = 20 + 5 = 25$

The standard error-pooled mean:

$MS_{\text{error-pooled}} = SS_{\text{error-pooled}} / \nu_{\text{error-pooled}} = 2.228 / 25 = 0.0891$

$$S_{\bar{X}_A - \bar{X}_B} = \sqrt{0.0891/6} = 0.122$$

3rd Find the probability (ρ) of the F-value:

Determine the number (a factor for the p value) of groups included in comparing each other:

a=5:

		#1	#2	#3	#4	#5	a
		3.800	2.333	2.183	1.967	1.533	
#1	3.800		1.467	1.617	1.833	2.267	5

a=4:

		#1	#2	#3	#4	#5	a
		3.800	2.333	2.183	1.967	1.533	
#1	3.800		1.467	1.617	1.833		4
#2	2.333			0.150	0.366	0.800	4

a=3:

		#1	#2	#3	#4	#5	a
		3.800	2.333	2.183	1.967	1.533	
#1	3.800		1.467	1.617			3
#2	2.333			0.150	0.366		3
#3	2.183				0.216	0.650	3

a=2:

		#1	#2	#3	#4	#5	a
		3.800	2.333	2.183	1.967	1.533	
#1	3.800		1.467				2
#2	2.333			0.150			2
#3	2.183				0.216		2
#4	1.967					0.434	2

The q-test method for comparing the each other (q-value list on the following table):

$$q = \frac{\left|\bar{X}_A - \bar{X}_B\right|}{S_{\bar{X}_A - \bar{X}_B}} = \frac{\left|\bar{X}_A - \bar{X}_B\right|}{\sqrt{\frac{MS_{error}}{2} \times \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

A with B	$\bar{X}_A - \bar{X}_B$	# of a	q value	p=0.05	p=0.01	P
(1)	(2)	(3)	(4)	(5)	(6)	(7)

1 w. 5	2.267	5	18.582	4.23	5.29	√
1 w. 4	1.833	4	15.025	3.96	5.02	√
1 w. 3	1.617	3	13.254	3.58	4.64	√
1 w. 2	1.467	2	12.025	2.95	4.02	√
2 w. 5	0.800	4	6.557	3.96	5.02	√
2 w. 4	0.366	3	3.000	3.58	4.64	
2 w. 3	0.150	2	1.230	2.95	4.02	
3 w. 5	0.650	3	5.328	3.58	4.64	√
3 w. 4	0.210	2	1.770	2.95	4.02	
4 w. 5	0.434	2	3.557	2.95	4.02	√

The mark with "√ " indicated the significant meaning of comparing test. Note: the "V_{error}" (degree of freedom) is by using the closing 20 instead of 25 for finding the q-table value as the q-table only showed the value of the n=20.

The q-value and the probability (p) indicated that there are statistic significant differences between sample 1 with sample 5, sample 1 with sample 4, sample 1 with sample 3, sample 1 with sample 2, sample 2 with sample 5, sample 3 with sample 5, and sample 4 with sample 5. The others have no significant difference in their mean.

Chapter 6

Binomial Distribution

In probability theory and statistics, the binomial distribution with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes-no question, and each with its own boolean-valued outcome: success/yes/true/one (with probability p) or failure/no/false/zero (with probability q = 1 - p). A single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance.

The binomial distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution remains a good approximation, and is widely used.

Probability Mass Function

A probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value. Sometimes it is also known as the discrete density function. The probability mass function is often the primary means of defining a discrete probability distribution, and such functions exist for either scalar or multivariate random variables whose domain is discrete.

Permutations and Combinations

Number of permutations (order matters) of *n* things taker X at a time:	Number of combinations (order does not matter) of *n* things taker X at a time:
$P(n,\mathrm{x}) = \frac{n!}{(n-\mathrm{x})!}$	$C(n,\mathrm{x}) = \frac{n!}{(n-\mathrm{x})!\mathrm{x}!}$

A probability mass function differs from a probability density function (PDF) in that the latter is associated with continuous rather than discrete random variables. A PDF must be integrated over an interval to yield a probability, e.g., the price of a stock or ETF. PDFs are plotted on a graph typically resembling a bell curve, with the probability of the outcomes lying below the curve.

A typical PDF example:
Suppose bacteria of a certain species typically live 4 to 6 hours. The probability that a bacterium lives exactly 5 hours is equal to zero. A lot of bacteria live for approximately 5 hours, but there is no chance that any given bacterium dies at exactly 5.0000000000... hours. However, the probability that the bacterium dies between 5 hours and 5.01 hours is quantifiable. Suppose that the answer is 0.02 (i.e., 2%), then the probability that the bacterium dies between 5 hours and 5.001 hours might be about 0.002, since this time interval is one-tenth as long as the previous. Similarly, the probability that the bacterium dies between 5 hours and 5.0001 hours might be about 0.0002, and so on.

Unlike a probability, a probability density function can take on values greater than one; for example, the uniform distribution on the interval [0, ½] has probability density f (x) = 2 for $0 \leq x \leq 1/2$ and f (x) = 0 elsewhere.

The graph of a probability mass function:
All the values of this function must be non-negative and sum up to 1. The value of the random variable having the largest probability mass is called the mode, e.g. the mode of the following example is 0.512.

Suppose a biased coin comes up heads with probability 0.8 (π =0.8) when tossed. The all probability of 3 coins (n=3) in tosses is

$$[(1-\pi)+\pi]^n=(1-\pi)^n+\binom{n}{1}(1-\pi)^{n-1}\pi+\binom{n}{2}(1-\pi)^{n-2}\pi^2+\ldots+\pi^n$$

$$(0.2+0.8)^3=(0.2)+3(0.2)^2(0.8)+3(0.2)(0.8)^2+0.8^3$$

$$=0.008+0.096+0.384+0.512=1$$

Comparing with 0.008, 0.096 and 0.384, the probability of 0.512 is the mode.

Binomial Distribution Formulation

Probability mass function is the probability distribution of a discrete random variable, and provides the possible values and their associated probabilities. It is the function p: defined by

$$p(x)=\binom{n}{x}(1-\pi)^{n-x}\pi^x$$

$$\binom{n}{x}=\frac{n!}{X!\,(n-X)!}$$

π:probability **X = 0, 1, 2, …, n**

Example 1: a probability calculation

Suppose a biased coin comes up heads with probability 0.3 (π =0.3) when tossed. The probability of seeing exactly 4 heads (x=4) in 6 tosses (n=6) is

$$p(4) = \frac{6!}{4!\,(6-4)!}(1-0.3)^{6-4}(0.3)^{4}$$

$$= \frac{6x5x4x3x2x1}{4x3x2x1x2x1} \times 0.7^{2} \times 0.3^{4}$$

$$= 0.059535$$

Suppose a test-positive rate is 20% in a population, if you take 10 samples from this population, (1) what is the possibility of 8 samples being test-positive exactly? (2) what is the possibility of 1 sample being test-positive most likely? (3) what is the possibility of 8 samples being test-positive at least?

(1) $p(8) = [10!/8!(10-8)!] \times (0.8)^{2}(0.2)^{8} = 0.00007373$

(2) $p(1) = p(0) + p(1) = 0.8(0.8)^{9}+[10!/1!(10-1)!] \times (0.8)^{9}(0.2)^{8} = 0.3758$

(3) $Q(8) = p(8)+(p9)+(10) = 0.00007373+[10!/9!(10-9)!] \times (0.8)(0.2)^{9}+(0.2)(0.2)^{9}$
$= 0.00007793$

Poisson Approximation

The binomial distribution converges towards the Poisson distribution as the number of trials goes to infinity while the product *np* remains fixed or at least *p* tends to zero. Therefore, the Poisson distribution with parameter $\lambda = np$ can be used as an approximation to B(n, p) of the binomial distribution if *n* is sufficiently large and p is sufficiently small. According to two rules of thumb, this approximation is good if $n = 20$ and $p = 0.05$, or if $n = 100$ and $np = 10$.

Poisson binomial distribution
The binomial distribution is a special case of the Poisson binomial distribution, or general binomial distribution, which is the distribution of a sum of n independent non-identical Bernoulli trials B(pi).

Poisson Approximation to the Binomial
When the value of n in a binomial distribution is large and the value of p is very small, the binomial distribution can be approximated by a Poisson

distribution. If n > 20 and np < 5 OR nq < 5 then the Poisson is a good approximation.

Normal approximation
If n is large enough, then the skew of the distribution is not too great. In this case a reasonable approximation to B(n, p) is given by the normal distribution and this basic approximation can be improved in a simple way by using a suitable continuity correction. The basic approximation generally improves as n increases (at least 20) and is better when p is not near to 0 or 1.

Example 2: Mean and Standard Variance of binomial distribution

Suppose a biased coin comes up heads with probability 0.8 (π =0.8) when tossed. The all probability of 3 coins (n=3) in tosses is

$$(0.2+0.8)^3 = (0.2)+3(0.2)^2(0.8)+3(0.2)(0.8)^2+0.8^3$$

$$= 0.008+0.096+0.384+0.512 = 1$$

X	f=p(x)	fX	ΣfX
(1)	(2)	(3)	(4)
0	0.008	0	0
1	0.096	0.096	0.096
2	0.384	0.768	1.536
3	0.512	1.536	4.608
Sum	Σf=1	2.4000	6.2400

Note:
Column 3 (fX) = Column 1 x Column 2; Column 4 (ΣfX) = accumulation of fX.

The probability would be sorted as the table above and calculated by following formula:

$$\text{Mean } \mu = \frac{f\text{X}}{\Sigma f} = \frac{2.4000}{1} = 2.4$$

Standard Variance $\sigma = \sqrt{\frac{\Sigma f X^2 - (\Sigma f X)^2 / \Sigma f}{\Sigma f}} = \sqrt{\frac{6.24 - (2.4)^2/1}{1}} = 0.69$

$\mu = n\pi = 3 \times 0.8 = 2.4$

$\sigma = \sqrt{n\pi(1-\pi)} = \sqrt{3 \times 0.8 (1- 0.8)} = 0.69$

If using the rate (%), then

$\mu_p = \pi$

$\sigma_p = \sqrt{\pi(1-\pi)}$

When *n* is known, the parameter *p* can be estimated using the proportion of successes: $p=x/n$. This estimator is found using maximum likelihood estimator and also the method of moments. This estimator is unbiased and uniformly with minimum variance, proven using Lehmann-Scheff theorem, since it is based on a minimal sufficient and complete statistic. It is also consistent both in probability and in MS error.

If using the sample for the population rate (%), then

$$S_p = \sqrt{p(1-p) / n}$$

Applications

Statistical Inference (Estimation of parameters)

Confidence intervals (Estimation of Interval):

$$(p-\mu_a S_p,\ p+\mu_a S_p)$$

Example 3: Confidence interval

Suppose you have 8 test-negative results from 10 samples of a population, what is 95% confidence interval for this population?

n = 10, X = 8; Using x =(10-8)=2 and the percentage table of confidence interval, an interval estimate with a specific level of confidence - percentage table, to find the range is between 3 to 56; then calculating by 100-3=97 and 100-56=44, the 95% confidence interval for this population is 44—97%.

Example 4: Comparing of the two means

Sample A: a sample of 80, with test positive (+) 23, positive rate 28.75%;
Sample B: a sample of 85, with test positive (+) 13, positive rate 15.29%;
Q: is there the difference between two rates?

1st Making the null hypothesis and setting the test confidence level:

H_0: $p_1 = p_2$
H_1: $p_1 \neq p_2$
$\alpha = 0.05$

Formula:

$$u = \frac{p_1 - p_2}{\sqrt{p_c (1- p_c) \ (1/n_1 + 1/n_2)}}$$

where, p_1, p_2 are the probability of sample 1 and sample 2; p_c is the combined or pooled possibility of p_1 and p_2, calculating by

$$p_c = \frac{X_1+X_2}{n_1+n_2}$$

2nd To calculate it, do the following:

$n_1 = 80$, $X_1 = 23$, $p_1 = 0.2875$; $n_2 = 85$, $X_2 = 13$, $p_2 = 0.1529$;

$$p_c = \frac{X_1+X_2}{n_1+n_2} = \frac{23+13}{80+85} = 0.2182$$

$$u = \frac{p_1 - p_2}{\sqrt{p_c(1-p_c)(1/n_1 + 1/n_2)}} = \frac{0.2875-0.1529}{\sqrt{0.2182(1-0.2182)(1/80 + 1/85)}}$$

$$= 2.0921$$

3rd Find the probability (ρ) of the u-value:

From u-table, 2.0921 indicated the probability is: 0.05>p>0.02; with the confidence level, $\alpha = 0.05$; $\alpha > p$, reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Example 5: Comparing of more than two means

Before a treatment, take 3 samples from population and the observations are 38, 29 and 36 per litter from the samples; After a treatment, take two samples from the same population and the observations are 25 and 18 per little, is there a difference?

1st Making the null hypothesis and setting the test confidence level:

$\boldsymbol{H_0}$: $p_1 = p_2$
$\boldsymbol{H_1}$: $p_1 \neq p_2$
$\boldsymbol{\alpha = 0.05}$

2nd To calculate it, do the following:

$\bar{X}_1 = (38+29+36)/3 = 34.33$, $n_1=3$;

$\bar{X}_2 = (25+18)/2 = 21.50$, $n_2 =2$;

$$u = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\bar{X}_1}{n_1} + \frac{\bar{X}_2}{n_2}}} = \frac{34.33 - 21.50}{\sqrt{\frac{34.33}{3} + \frac{21.50}{2}}}$$

$$= 2.723$$

3rd Find the probability (ρ) of the u-value:

From u-table, 2.723 indicated the probability is: 0.01>p>0.005; with the confidence level, $\alpha = 0.05$; $\alpha > p$, reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Note: whenever the normal distributions or approximation to the normal distributions are considered with the same sampling numbers, the simple formula for u-test could be applied:

$$u = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\bar{X}_1 + \bar{X}_2}} = \frac{\Sigma\bar{X}_1 - \Sigma\bar{X}_2}{\sqrt{\Sigma\bar{X}_1 + \Sigma\bar{X}_2}}$$

Chapter 7

Chi-squared Test

A chi-squared test, also written as X^2 test, is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof. Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

In the standard applications of this test, the observations are classified into mutually exclusive classes. If the null hypothesis that there are no differences between the classes in the population is true, the test statistic computed from the observations follows a X^2 frequency distribution. The purpose of the test is to evaluate how likely the observed frequencies would be assuming the null hypothesis is true.

Test statistics that follow a X^2 distribution occurs when the observations are independent and normally distributed, which assumptions are often justified under the central limit theorem. There are also X^2 tests for testing the null hypothesis of independence of a pair of random variables based on observations of the pairs.

Chi-squared tests often refers to tests for which the distribution of the test statistic approaches the X^2 distribution asymptotically, meaning that the sampling distribution (if the null hypothesis is true) of the test statistic approximates a chi-squared distribution more and more closely as sample sizes increase.

At the end of 19th century, Pearson noticed the existence of significant skewness within some biological observations. In order to model the observations regardless of being normal or skewed, Pearson, in a series of articles published from 1893 to 1916 devised the Pearson distribution, a family of continuous probability distributions, which includes the normal distribution and many skewed distributions, and proposed a method of

statistical analysis consisting of using the Pearson distribution to model the observation and performing a test of goodness of fit to determine how well the model really fits to the observations.

Applications

In cryptanalysis, the chi-squared test is used to compare the distribution of plaintext and (possibly) decrypted ciphertext. The lowest value of the test means that the decryption was successful with high probability. This method can be generalized for solving modern cryptographic problems.

In bioinformatics, chi-squared test is used to compare the distribution of certain properties of genes (e.g., genomic content, mutation rate, interaction network clustering, etc.) belonging to different categories (e.g., disease genes, essential genes, genes on a certain chromosome etc.).

In general:
If the calculated χ^2 > critical value - reject H_0 hypothesis and accept H_1.
If the calculated χ^2 < critical value - do not reject H_0.

Example 1: chi-squared test for categorical data

Contingency tables are used to determine whether 2 distinct variables are linked. To be able to quantify such linkage, we use the chi-squared (X^2) test.

Individual members of the sample/population are assigned to the appropriate cell of the contingency table according to their values for the two variables. When the table has only two rows or two columns this is equivalent to the comparison of proportions. In this case it is called four-fold table.

The use of the chi-squared test is not confined to nominal and ordinal data but can also be used for continuous variables that have been categorized. The procedure described for four-fold table can be easily applied for any contingency table, or a fourfold table.

The variables can be: Qualitative, Discrete quantitative and Continuous quantitative, whose values have been grouped (i.e. intervals).

When there are two such variables the data are arranged in a contingency table: Variable #1 -> rows Variable #2 -> columns

Test a two samples' mean by rate in a fourfold table:

Treatments	Positive (+) finding	Negative (-) finding	Total	Positive Rate (%)	Negative Rate (%)
A Treatment for A group	52(57.18)	19(13.82)	71	73.24	26.76
B Treatment for B group	39(33.82)	3(8.18)	42	92.86	7.14
Total	91	22	113	80.53	19.46

where, the number of finding, 52, 19, 39 and 3 are from the "raw" data of the actual frequency by a fourfold table; the total number are sum of them respectively; the rate of: 73.24%=(52/71)x100%, 92.86%=(39/42)x100%, 80.53%=(91/113)x100%.

While, (57.18)=71x80.53%, (13.82)=71x19.46%; (33.83)=42x80.53%, (8.13)=42x19.46%. They are the theoretical frequency.

1st Making the null hypothesis and setting the test confidence level:

H_0: $p_1 = p_2$
H_1: $p_1 \neq p_2$
$\alpha = 0.05$

2nd The value to be tested by (X^2) test is based on the formula as follows:

$$\chi^2 = \Sigma \frac{(A - T)^2}{T}$$

where, A is actual frequency; T is theoretical frequency, based on the hypnosis of *Ho* – the rate with no significant difference.

$$\chi^2 = \Sigma \frac{(A - T)^2}{T} = \frac{(52-57.18)^2}{57.18} + \frac{(19-13.82)^2}{13.82} + \frac{(39-33.82)^2}{33.82} + \frac{(3-8.18)^2}{8.18}$$

$$= 6.48$$

In order to interpret this chi-squared statistic, we need to know the number of degrees of freedom (df) involved For a contingency table this is given in general by the formula df = (number of rows - 1) x (number of columns - 1) or

v = (number of rows –1)(number of columns –1) = (2-1)(2-1) = 1

3rd Find the probability (ρ) of the χ^2-value:

From the χ^2 –table and v=1, find that 0.025>p>0.01, with $\alpha = 0.05$, reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Alternative method for the calculation

For the paired data, the χ^2 –table could be also expressed as following type and then applied in an alternative formula to calculate it.

Treatments	Positive (+) finding	Negative (-) finding	Total
A Treatment for A group	52(a)	19(b)	71(a+b)
B Treatment for B group	39(c)	3(d)	42(c+d)
Total	91(a+c)	22(b+d)	113(n)

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} = \frac{(52x3-19x39)^2 x113}{71x42x91x22} = 6.48$$

We can find that the same χ^2-value is calculated from the alternative formula. The method omits the calculation of the theological frequencies (or rates) in the table and therefore is a simple solution for this special 2x2 table.

Correction for Continuity

Chi-square is calculated only if all expected cell frequencies are equal to or greater than 5. The Yates value is corrected for continuity; the Pearson value is not. Both probability estimates are non-directional.

Yates's correction for continuity
In statistics, Yates's correction for continuity (or Yates's chi-squared test) is used in certain situations when testing for independence in a contingency table. It aims at correcting the error introduced by assuming that the discrete probabilities of frequencies in the table can be approximated by a continuous distribution (chi-squared). In some cases, Yates's correction may adjust too far, and so its current use is limited.

Correction for approximation error
Using the chi-squared distribution to interpret Pearson's chi-squared statistic requires one to assume that the discrete probability of observed binomial frequencies in the table can be approximated by the continuous chi-squared distribution. This assumption is not quite correct, and introduces some error.

To reduce the error in approximation, Frank Yates, an English statistician, suggested a correction for continuity that adjusts the formula for Pearson's chi-squared test by subtracting 0.5 from the difference between each observed value and its expected value in a 2x2 contingency table. This reduces the chi-squared value obtained and thus increases its p-value.

The effect of Yates's correction is to prevent overestimation of statistical significance for small data. This formula is chiefly used when at least one cell of the table has an expected count smaller than 5. Unfortunately, Yates's correction may tend to overcorrect. This can result in an overly conservative result that fails to reject the null hypothesis when it should (a type II error).

(1) $1<T<5$, $n>40$,

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

The following is Yates's corrected version of Pearson's chi-squared statistics:

(2) T<1 or n<40,

$$\chi^2 = \sum \frac{(|A - T| - 0.5)^2}{T}$$

$$\chi^2 = \frac{(|ad - bc| - n/2)^2 \, n}{(a+b)(c+d)(a+c)(b+d)}$$

Contingency table (RxC table)

$$\chi^2 = n \left(\sum \frac{A^2}{n_1 n_2} - 1 \right)$$

Example 2: Correction for approximation error

Suppose we have a set of data as follows:

	I	II	III	IV	Total
A	50	48	18	72	188
B	105	10	7	23	145
Total	155	58	25	95	333

1st Making the null hypothesis and setting the test confidence level:

H_0: $R_1 = R_2$
H_1: $R_1 \neq R_2$
$\alpha = 0.05$

2nd The value to be tested by (X^2) test is based on the formula as follows:

$$\chi^2 = 333\left(\frac{50^2}{188x155} + \frac{48^2}{188x58} + \frac{18^2}{188x25} + \frac{72^2}{188x95} + \frac{105^2}{145x155} + \frac{10^2}{145x58} + \frac{7^2}{145x25} + \frac{23^2}{145x95} - 1\right)$$

$$= 70.143$$

$$v = (2-1)(4-1) = 3$$

3rd Find the probability (ρ) of the χ^2-value:

From the χ^2 –table and v=1, we find 0.005>p, with $\alpha = 0.05$, then we reject the H_0 and accept H_1. In this example, we reject the null hypothesis, meaning: there is association between **R**1 and **R**2 and this conclusion has less than 5% probability that there could be huge differences in the observed values arising just by chance.

Exact Probabilities in 2x2 Table

Fisher Exact Probability Test

Logic and Procedure:

Consider a 2x2 contingency table of the sort described above, with the cell frequencies represented by a, b, c, d, and the marginal totals represented by a+b, c+d, a+c, b+d, and n.

	+	-	Totals
+	a	b	a+b
-	c	d	c+d
Totals	a+c	b+d	

If there were no systematic association between the variables A and B within the population from which the cell frequencies are randomly drawn, the

probability of any particular possible array of cell frequencies, a, b, c, d, given fixed values for the marginal totals a+b, c+d, etc., would be given by the hypergeometric rule

$$\frac{\frac{(a+c)!}{a!c!} \times \frac{(b+d)!}{b!d!}}{\frac{n!}{(a+b)!(c+d)!}}$$

which for computational purposes reduces to

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Also, the degree of disproportion within any array of cell frequencies—in effect, the degree of ostensible association between variables A and B within the sample—can be measured by the absolute difference

$$\left| \frac{a}{a+b} - \frac{c}{c+d} \right|$$

Example 3: Fisher Exact Probability Test

Suppose we have a data, 28 samples, from a marched-study as follows:

	+	-	Total
+	11(a)	9(b)	20
-	1(c)	7(d)	8
Total	12	16	28

1st Making the null hypothesis and setting the test confidence level:

H_0: $R_a = R_d$
H_1: $R_a \neq R_d$
$\alpha = 0.05$

2nd The value to be tested by (X^2) test is based on the formula as follows:

$$\chi^2 = \frac{(|ad - bc| - n/2)^2\, n}{(a+b)(c+d)(a+c)(b+d)} = \frac{(|11x7-9x1|-28/2)^2 x28}{20x88x12x16} = 2.66$$

$v = 1$

3rd Find the probability (ρ) of the χ^2-value:

From the χ^2 –table, $\chi^2 = 2.66$ and v=1, we find 0.25>p>0.10, with $\alpha = 0.05$, then we can't reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_0}$. In this example, the null hypothesis, meaning: there is no statistic significant association between **R**a and **R**d and this conclusion has more than 95% probability that there would be true.

Example 4: Further to test "b" and "c"

1st Making the null hypothesis and setting the test confidence level:

$\boldsymbol{H_0}$: $R_b = R_c$
$\boldsymbol{H_1}$: $R_b \neq R_c$
$\boldsymbol{\alpha = 0.05}$

2nd The value to be tested by (X^2) test is based on the formula as follows:

$$\chi^2 = \frac{(|b- c| -1)^2}{b+c} = \frac{(|9- 1| -1)^2}{9+1} = 4.90$$

3rd Find the probability (ρ) of the χ^2-value:

From the χ^2 –table, $\chi^2 = 2.66$ and v=1, we find 0.05>p>0.0025, with $\alpha = 0.05$, then we reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$. In this example, we reject the null hypothesis, meaning: there is association between **R**b and **R**c and this conclusion has less than 5% probability that there could be huge differences in the observed values arising just by chance.

Goodness-of-fit Test

The goodness of fit of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions (see Kolmogorov-Smirnov test), or whether outcome frequencies follow a specified distribution (see Pearson's chi-squared test). In the analysis of variance, one of the components into which the variance is partitioned may be a lack-of-fit sum of squares.

Pearson's chi-squared test
Pearson's chi-squared test uses a measure of goodness of fit which is the sum of differences between observed and expected outcome frequencies (that is, counts of observations), each squared and divided by the expectation:

Chapter 8

Nonparametric Statistic Analysis

Parametric Statistics and Nonparametric Statistics

What is the difference between a parametric and a nonparametric test?

Parametric tests assume underlying statistical distributions in the data. Therefore, several conditions of validity must be met so that the result of a parametric test is reliable. For example, Student's t-test for two independent samples is reliable only if each sample follows a normal distribution and if sample variances are homogeneous.

The advantage of using a parametric test instead of a nonparametric equivalent is that the former will have more statistical power than the latter. In other words, a parametric test is more able to lead to a rejection of H0. Most of the time, the p-value associated to a parametric test will be lower than the p-value associated to a nonparametric equivalent that is run on the same data. Parametric tests often have nonparametric equivalents. You will find different parametric tests with their equivalents when they exist in this grid.

Nonparametric tests are more robust than parametric tests. In other words, they are valid in a broader range of situations (fewer conditions of validity). Nonparametric tests do not rely on any distribution. They can thus be applied even if parametric conditions of validity are not met.

Nonparametric statistics is the branch of statistics that is not based solely on parametrized families of probability distributions. Nonparametric statistics is based on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified. Nonparametric statistics includes both descriptive statistics and statistical inference. Nonparametric tests are often used when the assumptions of parametric tests are violated.

Nonparametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of nonparametric methods may be necessary when data have a ranking

but no clear numerical interpretation, such as when assessing preferences. In terms of levels of measurement, non-parametric methods result in ordinal data.

As nonparametric methods make fewer assumptions, their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Also, due to the reliance on fewer assumptions, non-parametric methods are more robust.

Another justification for the use of nonparametric methods is simplicity. In certain cases, even when the use of parametric methods is justified, nonparametric methods may be easier to use. Due both to this simplicity and to their greater robustness, nonparametric methods are seen by some statisticians as leaving less room for improper use.

The wider applicability and increased robustness of nonparametric tests comes at a cost: in cases where a parametric test would be appropriate, nonparametric tests have less power. In other words, a larger sample size can be required to draw conclusions with the same degree of confidence.

Nonparametric Models

Nonparametric models differ from parametric models in that the model structure is not specified a priori but is instead determined from data. The term nonparametric is not meant to imply that such models completely lack parameters but that the number and nature of the parameters are flexible and not fixed in advance.

For instance, a histogram is a simple nonparametric estimate of a probability distribution; Data envelopment analysis provides efficiency coefficients similar to those obtained by multivariate analysis without any distributional assumption.

Methods

Nonparametric (or distribution-free) inferential statistical methods are mathematical procedures for statistical hypothesis testing which, unlike parametric statistics, make no assumptions about the probability

distributions of the variables being assessed. Order statistics, which are based on the ranks of observations, is one example of such statistics.

Signed Rank Test

The Wilcoxon signed-rank test is a nonparametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test (also known as "t-test for matched pairs" or "t-test for dependent samples") when the distribution of the difference between two samples' means cannot be assumed to be normally distributed. A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution.

Assumptions

Data are paired and come from the same population. Each pair is chosen randomly and independently. The data are measured on at least an interval scale when, as is usual, within-pair differences are calculated to perform the test (though it does suffice that within-pair comparisons are on an ordinal scale)

Example 1: Compare the means of paired data

Suppose a set of 12 samples, before treatment and after treatment, as follows, Question: is there difference by treatment?

				Rank	
n	Before	After	Difference (D)	+	-
(1)	(2)	(3)	(4)	(5)	(6)
1	76	93	-27		10
2	71	68	3	1	
3	70	65	5	4	
4	61	65	-4		3
5	80	93	-13		9
6	59	78	-19		12
7	74	83	-9		8
8	62	79	-17		11
9	79	98	-9		7

10	72	78	-6		5
11	84	90	-6		6
12	63	60	3	2	
n=12			Total	7	71

1st Making the null hypothesis and setting the test confidence level:

H_0: $M = 0$ (M: median. There is no difference by treatment)
H_1: $M \neq 0$ (There is a difference by the treatment)
$\alpha = 0.05$

2nd Calculate the difference and list on the table on column (5) and (6)

If the absolution number is same, but the positive and negative difference, then use average of them for rank.
If they are same value, for example of sampling 10 and 11, both of "-6" could be ranked as order in 5 and 6 on column (5) and (6) respectively.

Use the smaller one of ranking number as the "T", the total of 7 in column (5) in this example.

3rd Find the p-value

From T-table with n=12 and T=7, we find p=0.01, $\alpha = 0.05$, and reject the H_0 and accept H_1.

The close to u-test

When n>50, use following formula:

$$u = \frac{| T-n(n+1)/4 | - 0.5}{\sqrt{n(n+1)(2n-1)/24}}$$

where, “T” value is close to mean of “n(n+1)/4”; Square Variance is close to “n(n+1)(2n-1)/24; “0.5” is adjustment of correction for continuous. Then, the formula is close to a normal distribution.

When there are many of the same value on ranking in test, for example the total of same rank over 50, then use following formula for correction.

$$u = \frac{|\ T-n(n+1)/4\ | - 0.5}{\sqrt{\quad n(n+1)(2n-1)/24 - \Sigma(t_j^3-t_j)/48}}$$

where, t_j is number of the same difference. For example, if there are 2 of 3, 2 of 6, 2 of 17, the tj would be: $t_1=2$, $t_2=2$, $t_3=2$.

Example 2: The multiple marched cases

The multiple marched cases as follows:

	A	Rank	B	Rank	C	Rank	
Feb.	11.4	(3)	5.8	(2)	3.5	(1)	
Apr.	6.4	(1)	8.6	(3)	7.5	(2)	
Jun.	11.2	(3)	7.0	(1)	9.8	(2)	
Aug.	13.8	(3)	10.8	(2)	10.4	(1)	
Oct.	7.3	(1)	8.8	(2)	9.3	(3)	
Dec.	8.3	(3)	6.2	(2)	2.5	(1)	
Total (Ri)	14		12		10		
Average (R)	12		12		12		
							Sum(M)
$(Ri-R)^2$	4		0		4		8

Note:
A, B and C are ranked within the same month (the same row in table).
The average (R) is calculated: (14+12+10)/3=12.

1st Making the null hypothesis and setting the test confidence level:

H_0: $M_A = M_B = M_C = 0$
H_1: $M_A \neq M_B \neq M_C \neq 0$
$\alpha = 0.05$

2nd To calculate it, do the following:

Calculate the average R: (14+12+10)/3=12

Calculate the Square Variance $(R_i-R)^2$: $(14-12)^2=4$; $(12-12)^2=0$; $(10-12)^2=4$;

M-value: 4+0+4=8

3rd Determine the p-value:

From M-table with n=6 and k=3, we find $M_{0.05}=42$; As M-value =8 and less than 42, p>0.05; with $\alpha = 0.05$, we can't reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_0}$. Thus, we can't believe that there are differences among the A, B and C.

Example 3: Test of rank data

Suppose we have two sample cases as follows:

A Rank	B Rank
5(1)	17(9)
5(2)	18(10.5)
6(3)	20(12)
7(4)	25(14)
9(5)	34(15)
12(6)	43(16)
13(7)	44(17)
15(8)	
18(10.5)	
21(13)	

$n_A=10$ $T_A=59.5$ $n_B=7$ $T_B=93.5$

1st Making the null hypothesis and setting the test confidence level:

$\boldsymbol{H_0}$: $M_A = M_B$
$\boldsymbol{H_1}$: $M_A \neq M_B$
$\boldsymbol{\alpha = 0.05}$

2nd Rank the samples and calculate the T-value in the following table:

Test Index (1)	Two Samples A (2)	B (3)	Total (4)	Rank Range (5)	Rank Average (6)	Rank Totals A (7)	B (8)
<1	4		4	1 -- 4	2.5	10	
1 --	11		11	5 --15	10	110	
5 --	15	2	17	16 --32	24	360	48
10 --		10	10	33 --42	37.5		375
15 --		1	1	43	43		43
20 --		4	4	44 -- 47	45.5		182
25 --		2	2	48 -- 49	48.5		97
Total	n_A=30	n_B=19	49			T_A=480	T_B=745

Calculate rank total in Column (4);

Calculate rank range and accumulate the rank number in column (5).
For example: 1st row(<1) has a total of 4; 2nd row has a total of 11 and plus the 1st row of 4 to accumulate in total of 15, therefore the rank range would be from 5 to 15.

Calculate rank average.
For example: (1+4)/2=2.5; (5+15)/2=10; (16+32)/2=24; (33+42)/2=37.5; (43+43)/2=43; (44+47)/2=45.5; (48+49)/2=48.5.

Rank of total on column (7) = column (2) multiply the column (6).
For Example: 4x2.5=10 in 1st row.

Rank of total on column (8) = column (3) multiply the column (6).
For Example: 2x24=48 in 3rd row.

3rd Determine the p-value:

From T-table, $n_A < n_B$, T=93.5, p<0.005, $\alpha = 0.05$, reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Approximation to the Normal Distribution

Example 4: Approximation to the normal distribution method

Alternatively, when n1>20 and n2-n1>10, we can use the following formula to calculate the u-test value; when there are many of the same number in the same range of rank, we should use the adjusted u-test in this example:

$$u = \frac{| T_1 - n_1 (N+1)/2 | - 0.5}{\sqrt{\frac{n_1 n_2}{12N(N-1)} \left[N^3 - N - \Sigma(t_j^3 - t_j)\right]}}$$

$$= \frac{|745 - 19(49+1)/2 | - 0.5}{\sqrt{\frac{19(30)}{12(49)(49-1)} \left[(49)^3 - 49 - 7332\right]}} = 5.711$$

where, $\Sigma(t_j^3 - t_j) = (4^3-4)+(11^3-11)+(17^3-17)+(10^3-10)+(4^3-4)+(2^3-2)=7332$

From u-table, n=49, u=5.711, we find p<0.001, with $\alpha = 0.05$, we reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$. (The same result as previous test)

Multiply Samples (H-test)

H-test:

$$H = \frac{12}{N(N+1)} \Sigma \frac{R_i^2}{n_i} - 3(N+1)]$$

Correction for approximation error

When there are many of the same ranks, use the adjusted formula:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \Big/ \left[1 - \frac{\Sigma(t_j^3 - t_j)}{N-N}\right]$$

Example 5: Test for multiply samples

Suppose we have a set of data, Four Groups of A, B, C, and D, as follows:

	A	Rank	B	Rank	C	Rank	D	Rank	
	0.15	1	1.20	7.5	0.50	5.5	1.50	13	
	0.30	2	1.35	9	1.20	7.5	1.50	13	
	0.40	3	1.40	10.5	1.40	10.5	2.50	20	
	0.40	4	1.50	13	2.00	16	2.50	21	
	0.50	5.5	1.90	15	2.20	17			
			2.30	19	2.20	18			
Ri		15.5		74		74.5		67	
n_i		5		6		6		4	N=21

1st Making the null hypothesis and setting the test confidence level:

H_0: $M_A = M_B = M_C = M_D$
H_1: $M_A \neq M_B \neq M_C \neq M_D$
$\alpha = 0.05$

2nd Calculate rank of each group, do the following:

Two 0.40 in the same group of A, then, make their rank as ordinary way as 3 and4. However, two 0.50 in group A and group C, then, use (5+6)/2=5.5 (the average rank should be balanced as the same value in different groups)

Calculate Ri and n_i as showed in the table.

Total sample: N=5+6+6+4=21

Calculate H value.

$$H = \frac{12}{N(N+1)} \Sigma \frac{R_i^2}{n_i} - 3(N+1)\ ]$$

$$= \frac{12}{21(21+1)} [\ \frac{15.5^2}{5} + \frac{74^2}{6} + \frac{74.5^2}{6} + \frac{67^2}{4}\ ] - 3(21+1) = 12.13$$

Correction for approximation error
(If use the formula of correction for approximation error, the H value is 12.21, close to 12.13 above)

3rd Find the p-value

From the χ^2-table, k=4 and v=4-1=3, we find 0.01>P>0.005, with $\alpha = 0.05$, we reject the $\boldsymbol{H_0}$ and accept $\boldsymbol{H_1}$.

Example 6: Test for multiply rank data

Five Groups of A, B, C, D and E:

	A	B	C	D	E	Total	Rank Range	Rank Average
I	21	19				40	1 - 40	20.5
II	4	4	41	3		52	41-92	66.5
III		0	6	11	31	48	93-140	116.5
IV		2	3	15	42	62	141-202	171.5
V				21	77	98	203-300	251.5
ni	25	25	50	50	150	300		
Ri	696.5	998.5	3940	9335	30180			

1st Making the null hypothesis and setting the test confidence level:

$\boldsymbol{H_0}$: $M_A = M_B = M_C = M_D = M_E$
$\boldsymbol{H_1}$: $M_A \neq M_B \neq M_C \neq M_D \neq M_E$
$\boldsymbol{\alpha = 0.05}$

2nd Calculate rank range and rank average.

Calculate Ri and ni. (See the table for details)

N=25+25+50+50+150=300

Ri= Number in a group multiply the Rang Average respectively.
For example, Group A: 21x20.5+4x66.5=696.5.

Calculate H value:

As there are many of the same rank in the case, use the adjusted formula:

$$\Sigma(t_j^3-t_j) = (40^3-40)+(52^3-52)+(48^3-48)+(62^3-62)+(98^3-98) = 1494420$$

$$H = \frac{12}{N(N+1)} \Sigma \frac{R_i^2}{n_i} - 3(N+1) \Big/ [1 - \frac{\Sigma(t_j^3-t_j)}{N-N}]$$

$$= \frac{[\frac{12}{300(300+1)} \times (\frac{696.5^2}{25} + \frac{998.5^2}{25} + \frac{3940}{50} + \frac{9335^2}{50} + \frac{30180^2}{150}) - 3(300+1)]}{1494420 / (300^3-300)}$$

$$= 195.50$$

3rd Find the p value

From the χ^2-table, K=5, k=4, v=5-1=4, we find P<0.005; with $\alpha = 0.05$, we reject the H_0 and accept H_1.

Multiply Samples with Comparing Each of Two

$$t = \frac{R_A - R_B}{\sqrt{\frac{N(N+1)(N-1-H)}{12(N-k)} (\frac{1}{n_A} + \frac{1}{n_B})}}$$

v=N-k

Example 7: Compare each of two groups from the multiply samples

Based on case of Four Groups of A, B, C, and D before, the H value was 195.50.

	A	Rank	B	Rank	C	Rank	D	Rank	
	0.15	1	1.20	7.5	0.50	5.5	1.50	13	
	0.30	2	1.35	9	1.20	7.5	1.50	13	
	0.40	3	1.40	10.5	1.40	10.5	2.50	20	
	0.40	4	1.50	13	2.00	16	2.50	21	
	0.50	5.5	1.90	15	2.20	17			
			2.30	19	2.20	18			
Ri		15.5		74		74.5		67	
ni		5		6		6		4	N=21

Now take two groups from the 4 groups for testing.

Mean: R_{iA}=15.5/5=3.10; R_{iB}=74/6=12.33; R_{iC}=74.5/6=12.42; R_{iD}=67/4=16.75.
Difference: for example, $R_{iA} - R_{iB}$ = 3.10 – 12.33 = -9.23

$$t = \frac{R_A - R_B}{\sqrt{\frac{N(N+1)(N-1-H)}{12(N-k)}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

$$= \frac{3.10 - 12.33}{\sqrt{\frac{21(21+1)(21-1-12.21)}{12(21-4)}\left(\frac{1}{5} + \frac{1}{6}\right)}} = -3.629$$

The values could be sorted out as follows:

A v. B	n_A	n_B		t	p	Ho
1 v. 2	5	6	-9.23	-3.629	0.005>p>0.002	*x*
1 v. 3	5	6	-9.32	-3.664	0.002>p>0.001	*x*

1 v. 4	5	4	-13.65	-4.845	0.001>p	*x*
2 v. 3	6	6	-0.09	-0.038	p>0.50	√
2 v. 4	6	4	-4.42	-1.630	0.20>p>0.10	√
3 v. 4	6	4	-4.33	-1.597	0.20>p>0.10	√

From the calculation listed on the table above, we can determine the p and check the statistic significances on the H_o:
"1 vs. 2", "1 vs. 3" and "1 vs.4" indicated rejecting the H_o;
"2 vs. 3", "2 vs. 4" and "3 vs.4" indicated not rejecting the H_o.

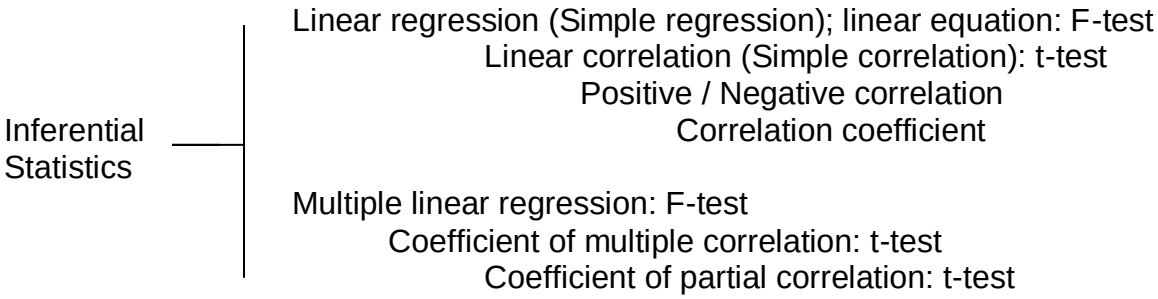
Chapter 9

Simple Regression Analysis

What Is Regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

The are two categories of regression analysis in inferential statistics and they can summarize up as follows:



Why is it called "linear" regression?

Linear implies the model functions along a straight or nearly straight line. Linear suggests that the relationship between dependent and independent variable can be expressed in a straight line. A Simple Linear Regression uses a single feature (independent variable) to predict a target (dependent variable) by fitting a best linear relationship, whereas Multiple Linear Regression uses more than one feature to predict a target variable by fitting a best linear relationship. In this chapter, we shall mainly focus Simple Linear Regression.

Simple Linear Regression

In statistics, simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable. The adjective simple refers to the fact that the outcome variable is related to a single predictor.

It is common to make the additional stipulation that the ordinary least squares (OLS) method should be used: the accuracy of each predicted value is measured by its squared residual (vertical distance between the point of the data set and the fitted line), and the goal is to make the sum of these squared deviations as small as possible. Other regression methods that can be used in place of ordinary least squares include least absolute deviations (minimizing the sum of absolute values of residuals) and the Theil-Sen estimator (which chooses a line whose slope is the median of the slopes determined by pairs of sample points). Deming regression (total least squares) also finds a line that fits a set of two-dimensional sample points, but (unlike ordinary least squares, least absolute deviations, and median slope regression) it is not really an instance of simple linear regression, because it does not separate the coordinates into one dependent and one independent variable and could potentially return a vertical line as its fit.

The remainder of the article assumes an ordinary least squares regression. In this case, the slope of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that the line passes through the center of mass (x, y) of the data points.

Fitting the regression line and the model function

$Y=a+bX$

which describes a line with slope “b” and y-intercept “a”. In general such a relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables; we call the unobserved deviations from the above equation the errors.

Suppose we observe n data pairs and call them {(Xi, Yi), i = 1, ... n}. We can describe the underlying relationship between Yi and Xi involving this error term Ei by

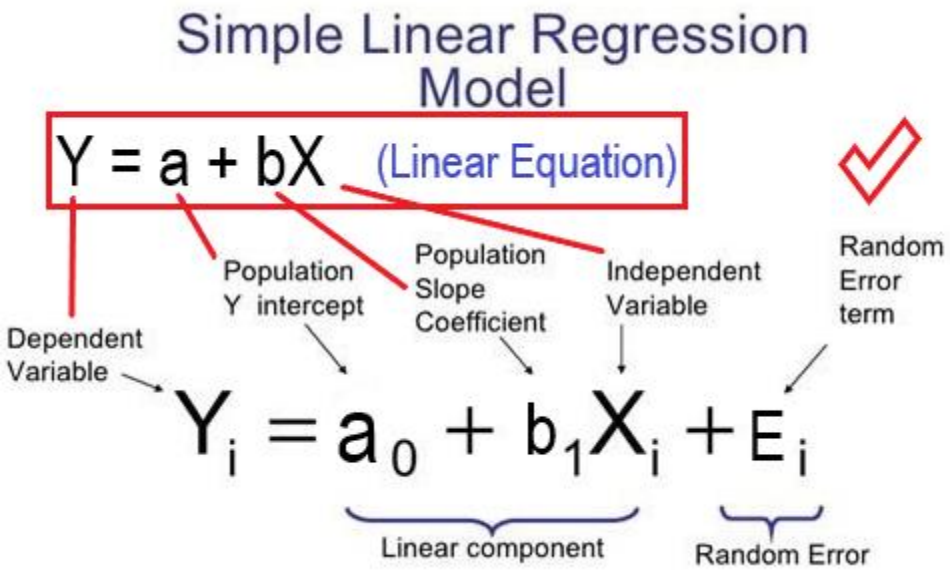
Yi=a+bXi+Ei

This relationship between the true (but unobserved) underlying parameters “a” and “b” and the data points is called a linear regression model.

In mathematical modeling, the dependent variable is studied to see if and how much it varies as the independent variables vary. In the simple stochastic linear model Yi=a+bXi+Ei the term Yi is the *i*-th value of the dependent variable and Xi is the *i*-th value of the independent variable. The term Ei is known as the "error" and contains the variability of the dependent variable not explained by the independent variable, and Ei=Yi-a-bXi.

The goal is to find estimated values “average a” and “average b” for the parameters “a” and “b” which would provide the "best" fit in some sense for the data points. As mentioned in the introduction, in this article the "best" fit will be understood as in the least-squares approach: a line that minimizes the sum of squared residuals “Ei” (differences between actual and predicted values of the dependent variable y), each of which is given by, for any candidate parameter values “a” and “b” (Ei=Yi-a-bXi)

The illustration of Linear equation and Linear regression model:



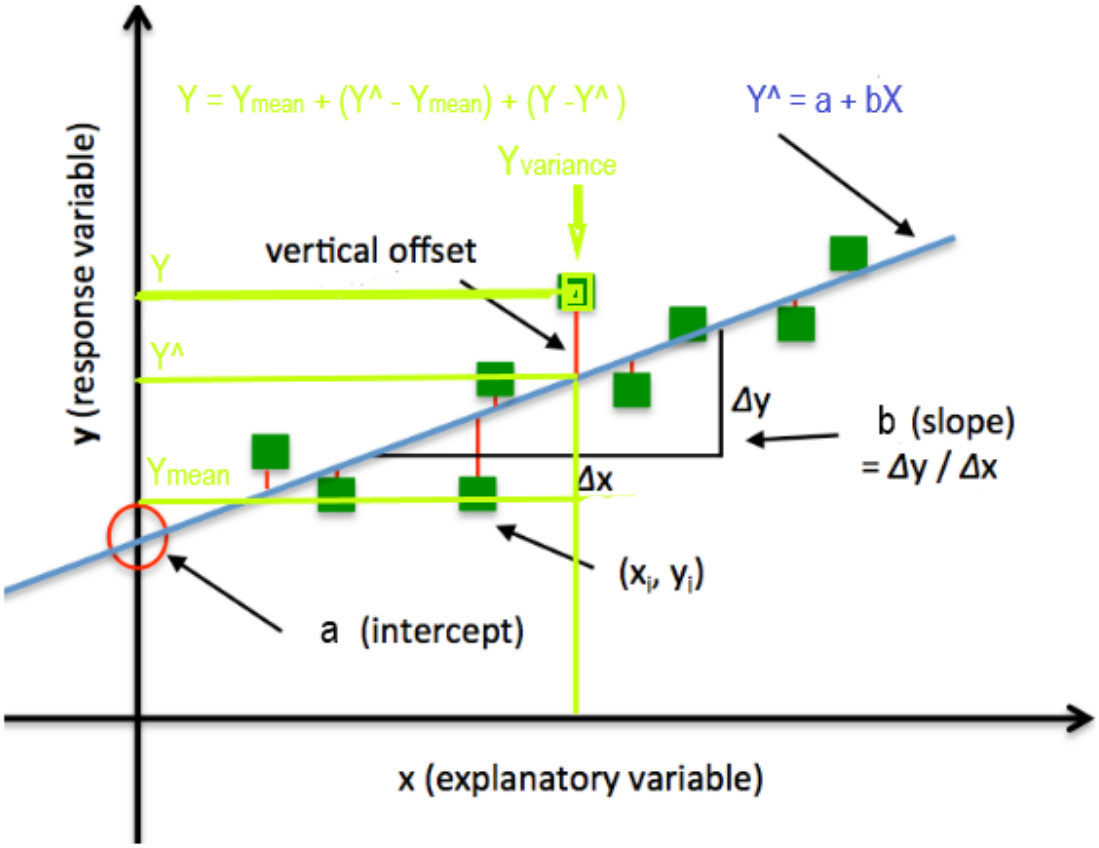
In mathematical least square method, "a" and "b" solve the following minimization problem:

$$b = \frac{\Sigma(X - X_m)(Y - Y_m)}{\Sigma(X - X_m)^2} = \frac{L_{xy}}{L_{xx}}$$

$$a = Y_m - bX_m$$

$$L_{xy} = \Sigma(X - X_m)(Y - Y_m) = \Sigma XY - (\Sigma X)(\Sigma Y)/n$$
$$L_{xx} = \Sigma(X - X_m)^2 = \Sigma X - (\Sigma X)^2/n$$
$$L_{yy} = \Sigma(Y - Y_m)^2 = \Sigma Y - (\Sigma Y)^2/n$$

The figure of Linear equation and Linear regression model:



The relationship of lighter green point (Y) with the linear equation ($y\hat{}=a+bX$) is supposed to be tested for the correlation.

$Y=Y_m+(Y\hat{}-Y_m)+(Y-Y\hat{}) \rightarrow (Y-Y_m) = (Y\hat{}-Y_m)+(Y-Y\hat{})$

If all the Y point value to be summed up, the total sum of square would be

$$\Sigma (Y - Y_m)^2 = \Sigma (Y\hat{}-Y_m)^2 + \Sigma (Y-Y\hat{})^2$$

$$SS_{total} = SS_{regression} + SS_{residual}$$

It indicated the relationship of sum of square:

$$SS_{total} = SS_{regression} + SS_{residual}$$

Where,
SS_{total}: Total sum of square, indicating the Y-variance.
$SS_{regression}$: Regression sum of square, indicating how X-variance to be correlation to the Y-variance.
$SS_{residual}$: Residual sum of square, indicating how other factors, not including X-variance, to be correlation to the Y-variance.

Degree of freedom:
$V_{total} = V_{regression} + V_{residual}$
$V_{total} = n-1$; $V_{regression} = 1$; $V_{residual} = n-2$;

Example 1: how to set up a simple linear equation

Suppose we have a set of data as follows:

X	1	2	3	4	5
Y	4.0	5.5	6.2	7.7	8.5

Calculating Slope and Intercepts

We all know from elementary geometry that equation of a straight line can be written as: $Y= a + bX$

1st Sort the table as follows:

	X (1)	X^2 (2)	Y (3)	Y^2 (4)	XY (5)
	1	1	4.0	16.00	4.0
	2	4	5.5	30.25	11.0
	3	9	6.2	38.44	18.6
	4	16	7.7	59.29	30.8
	5	25	8.5	72.25	42.5
Total	15	55	31.9	216.23	106.9
n	5		5		
Mean	3		6.38		

2nd Do the following calculation:

$L_{xy} = \Sigma(X - X_m)(Y - Y_m) = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 106.9-(15)(31.9)/5 = 11.20$
$L_{xx} = \Sigma(X - X_m)^2 = \Sigma X - (\Sigma X)^2/n = 55-15^2/5 = 10$

3rd Set up the Linear equation:

$$b = \frac{\Sigma(X - X_m)(Y - Y_m)}{\Sigma(X - X_m)^2} = \frac{L_{xy}}{L_{xx}} = \frac{11.20}{10} = 1.12$$

$a = Y_m - bX_m = 6.38 - 1.12x3 = 3.02$

The Linear equation: $Y = a + bX = 3.02 + 1.12X$

Further Statistic Consideration

When a simple linear equation is set up by the available sampling data and the linear equation is supposed to be existed, we basically have to answer two statistic questions:

1. is the linear regression equation right in statistics? - The answer for this question is to test the regression coefficient.
2. how can the differences in one variable be explained by the difference in a second variable (Linear correlation)? - The answer for this question is to test the correlation coefficient.

Test of a Regression Coefficient

Whether the linear equation is correlated or not, a statistic test should be conducted for finding the linear relationship between the variables (X) and (Y). In other words, the statistic testing the regression coefficient of the linear equation is needed for checking the variables (X) and (Y).

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.

The t-test is applicable to a variety of problems. In particular, it is applicable to the problem of testing the statistical significance of a regression coefficient. Under a set of assumptions that are usually referred to as the Gauss-Markov conditions, the t test can be used to test the significance of a regression coefficient.

In statistics, the Gauss-Markov theorem states that the ordinary least squares (OLS) estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance.

Regression Coefficient Test

How to determine a regression coefficient

Base on the linear equation set up in 3rd step above, we do the following regression coefficient test:

Use the linear equation to calculate the value in Column (6), (7) and (8) to fill out the following table.

	X (1)	X^2 (2)	Y (3)	Y^2 (4)	XY (5)	Y_{total} (6)	$Y-Y_{total}$ (7)	$(Y-Y_{total})^2$ (8)	$(Y-Y_{mean})^2$ (9)
	1	1	4.0	16.00	4.0	4.14	-0.14	0.0196	5.6644
	2	4	5.5	30.25	11.0	5.26	0.24	0.0576	0.7744
	3	9	6.2	38.44	18.6	6.38	-0.18	0.0324	0.0324
	4	16	7.7	59.29	30.8	7.50	0.20	0.0400	1.7424
	5	25	8.5	72.25	42.5	8.62	-0.12	0.0144	4.4944
Sum	15	55	31.9	216.23	106.9		0	0.1640	12.708
n	5		5						
Mean	3		6.38						

For example,

Column (6): $Y_{1\text{-total}} = a + bX_1 = 3.02 + 1.12x1 = 4.14$

Column (7): $Y_1 - Y_{1\text{-total}} = 4 - 4.14 = -0.14$

Column (8): $(Y_1 - Y_{1\text{-total}})^2 = (4 - 4.14)^2 = 0.0196$

Column (9): $(Y_1 - Y_{mean})^2 = (4 - 6.38)^2 = 5.6644$

Sum up the calculation:

The Linear equation: $Y = a + bX = 3.02 + 1.12X$

$$\Sigma (Y - Y_m)^2 = \Sigma (Y^\wedge - Y_m)^2 + \Sigma (Y - Y^\wedge)^2$$

$$SS_{total} = SS_{regression} + SS_{residual}$$

SS_{total}: $L_{yy} = \Sigma(Y - Y_m)^2 = \Sigma Y^2 - (\Sigma Y)^2/n = 216.23 - 31.9^2/5 = 12.708$
(Or, SS_{total}: $L_{yy} = \Sigma Y - (\Sigma Y)^2/n = 216.23 - 31.9^2/5 = 12.708$)

$L_{xy} = \Sigma(X - X_m)(Y - Y_m) = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 106.9 - (15)(31.9)/5 = 11.20$
$L_{xx} = \Sigma(X - X_m)^2 = \Sigma X^2 - (\Sigma X)^2/n = 55 - 15^2/5 = 10$

$SS_{residual} = \Sigma(Y - Y^\wedge)^2 = 0.1640$
(Or, $SS_{residual}$: $L^2_{xy} / L_{xx} = 11.2^{2/}/10 = 0.1640$
$L_{xy} = \Sigma(X - X_m)(Y - Y_m) = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 106.9 - (15)(31.9)/5 = 11.20$
$L_{xx} = \Sigma(X - X_m)^2 = \Sigma X^2 - (\Sigma X)^2/n = 55 - 15^2/5 = 10$)

$SS_{regression} = SS_{total} - SS_{residual} = 12.708 - 0.1640 = 12.544$

$V_{total} = V_{regression} + V_{residual}$ => $V_{total} = n-1$; $V_{regression} = 1$; $V_{residual} = n-2$;

Example 2: Methods of F-test to test a regression coefficient

1st Making the null hypothesis and setting the test confidence level:

H_0: $\beta = 0$, (No linear regression)
H_1: $\beta \neq 0$,
$\alpha = 0.05$

2nd Calculate the F value according to the following formula:

$$F = \frac{MS_{regression}}{MS_{residual}} = \frac{SS_{regression} / V_{regression}}{SS_{residual} / V_{residual}} = \frac{12.544 / 1}{0.164 / (5-2)} = 229.32$$

3rd Find the p value

From the F-table, with $V_{regression} = 1$ and $V_{residual} = 3$, we find $p<0.01$; with $\alpha = 0.05$ and $\alpha > p$, we have statistic significant confidence, with 95%, accepting the effects of X-variable correlated with Y-variable in the linear regression mode.

Example 3: Methods of t-test to test a regression coefficient

1st Making the null hypothesis and setting the test confidence level:

H_0: $\beta = 0$, (No linear regression)
H_1: $\beta \neq 0$,
$\alpha = 0.05$

2nd Calculate the t value according to the following formula:

$$t = \frac{b-0}{} = \frac{b}{}$$

$$S_b \qquad S_{yx} \;/\; \sqrt{L_{xx}}$$

Where, S_b, standard error of regression coefficient b; S_{yx}, standard deviation of Y for fixed X.

The Linear equation: $Y = a + bX = 3.02 + 1.12X$; $a=3.02$; $b=1.12$.

$$S_{yx} = \sqrt{\frac{\Sigma(Y - Y\hat{})^2}{n-2}} = \sqrt{\frac{SS_{residual}}{n-2}} = \sqrt{\frac{0.164}{5-2}} = 0.2338$$

$$t = \frac{b}{S_{yx} \;/\; \sqrt{L_{xx}}} = \frac{1.12}{0.2338 \;/\; \sqrt{10}} = 15.149$$

3rd Find the p value

From the t-table, with $V_{regression} = 1$ and $V_{residual} = 3$, we find $p<0.01$; with $\alpha = 0.05$ and $\alpha > p$, we reject the $\boldsymbol{H_0}$ and have statistic significant confidence, with 95%, accepting the effects of X-variable correlated with Y-variable in the linear regression mode.

Applications of Simple Linear Regression

1. Description the regression of the two variance, measurable amount.
2. Forecast: X as forecast facto and linear equation as formula to make the forecast of reliable interval. Syx could be used as the index for regression.
3. Statistical control: based on the regression equation and knowing the variance of Y, then, it could conduct the control on variance of X under some point of limitation (by t-table value).

For instance:

Linear Regression is a very powerful statistical technique and can be used to generate insights on consumer behaviour, understanding business and factors influencing profitability. Linear regressions can be used in business to

evaluate trends and make estimates or forecasts. For example, if a company's sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months.

Linear regression can also be used to analyze the marketing effectiveness, pricing and promotions on sales of a product. For instance, if company XYZ, wants to know if the funds that they have invested in marketing a particular brand has given them substantial return on investment, they can use linear regression. The beauty of linear regression is that it enables us to capture the isolated impacts of each of the marketing campaigns along with controlling the factors that could influence the sales. In real life scenarios there are multiple advertising campaigns that run during the same time period. Supposing two campaigns are run on TV and Radio in parallel, a linear regression can capture the isolated as well as the combined impact of running this ads together.

Linear regression can be also used to assess risk in financial services or insurance domain. For example, a car insurance company might conduct a linear regression to come up with a suggested premium table using predicted claims to Insured Declared Value ratio. The risk can be assessed based on the attributes of the car, driver information or demographics. The results of such an analysis might guide important business decisions.

In the credit card industry, a financial company maybe interested in minimizing the risk portfolio and wants to understand the top five factors that cause a customer to default. Based on the results the company could implement specific EMI options so as to minimize default among risky customers.

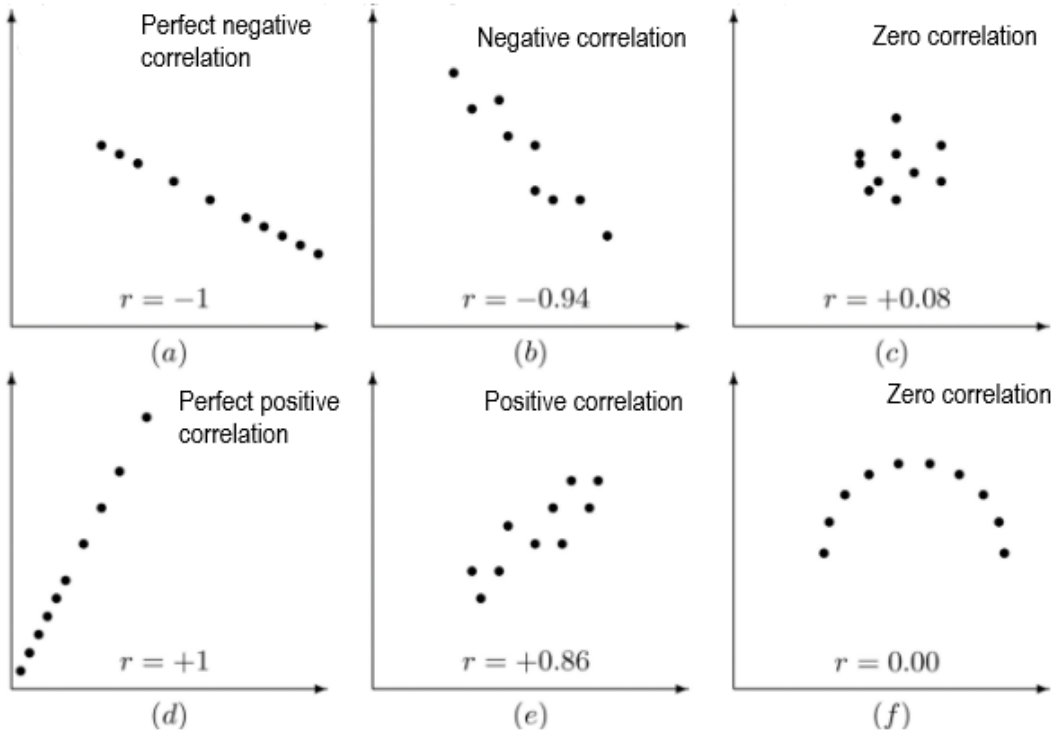
While linear regression has limited applicability in business situations because it can work only when the dependent variable is of continuous nature, it still is a very well known technique in the situations it can be used. It assumes a linear relation between the independent and dependent variables. It must be noted that sometimes transformations can also be applied to non-linear relationships to make them applicable in a linear regression model.

Linear Correlation

Simple correlation can be used for bivariate normal distribution and indicated:

Zero correlation
Positive correlation
Negative correlation
Perfect negative correlation
Perfect positive correlation

The following figures have showed how the typical correlations are expressed by “r” value between -1.0 to 1.0:



Understanding the Correlation Coefficient

The correlation coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference

in a second variable, when predicting the outcome of a given event. In other words, this coefficient, which is more commonly known as R-squared (or R^2), assesses how strong the linear relationship is between two variables, and is heavily relied on by researchers when conducting trend analysis. To cite an example of its application, this coefficient may contemplate the following question: if a woman becomes pregnant on a certain day, what is the likelihood that she would deliver her baby on a particular data in the future? In this scenario, this metric aims to calculate the correlation between two related events: conception and birth.

The correlation coefficient of determination is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. This correlation, known as the "goodness of fit," is represented as a value between 0.0 and 1.0. A value of 1.0 indicates a perfect fit, and is thus a highly reliable model for future forecasts, while a value of 0.0 would indicate that the calculation fails to accurately model the data at all. But a value of 0.20, for example, suggests that 20% of the dependent variable is predicted by the independent variable, while a value of 0.50 suggests that 50% of the dependent variable is predicted by the independent variable, and so forth.

Graphing the Correlation Coefficient

On a graph, the goodness of fit measures the distance between a fitted line and all of the data points that are scattered throughout the diagram. The tight set of data will have a regression line that's close to the points and have a high level of fit, meaning that the distance between the line and the data is small. Although a good fit has an R^2 close to 1.0, this number alone cannot determine whether the data points or predictions are biased. It also doesn't tell analysts whether the coefficient of determination value is intrinsically good or bad. It is at the discretion of the user to evaluate the meaning of this correlation, and how it may be applied in the context of future trend analyses.

Types of "goodness of fit"

There are several different measures for the degree of correlation in data, depending on the kind of data: principally whether the data is a measurement, ordinal, or categorical.

Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics. They all assume values in the range from -1 to +1, where ±1 indicates the strongest possible agreement and 0 the strongest possible disagreement.

Pearson product-moment correlation coefficient

The Pearson product-moment correlation coefficient, also known as r, R, or Pearson's r, is a measure of the strength and direction of the linear relationship between two variables that is defined as the covariance of the variables divided by the product of their standard deviations. This is the best-known and most commonly used type of correlation coefficient. When the term "correlation coefficient" is used without further qualification, it usually refers to the Pearson product-moment correlation coefficient.

$$r = \frac{\Sigma(X - X_m)\,\Sigma(Y - Y_m)}{\sqrt{\Sigma(X - X_m)^2\,\Sigma(Y - Y_m)^2}} = \frac{Lxy}{\sqrt{Lxx\,Lyy}}$$

r could be the value between –1 and +1 to indicate the correlation. r-table to find the p value.

Example 4: Calculation of correlation coefficient of determination

Suppose we have the data as follows:

N	X	X^2	Y	Y^2	X*Y
1	77	5929	87	7569	6699
2	78	6084	90	8100	7020
3	79	6241	89	7921	7031
4	80	6400	90	8100	7200
5	81	6561	91	8281	7371
6	82	6724	89	7921	7298
7	83	6889	91	8281	7553
8	84	7056	92	8464	7728
9	76	5776	86	7396	6536
10	79	6241	88	7744	6952
Sum (Σ)	799	63901	893	79777	71388

1st Summary of calculation:

$\Sigma X = 799$; $\Sigma X^2 = 63901$; $\Sigma Y = 893$; $\Sigma Y^2 = 79777$; $\Sigma XY = 71388$.

$L_{xx} = \Sigma X^2 - (\Sigma X)^2/n = 63901 - 799^2/10 = 60.9$
$L_{yy} = \Sigma Y^2 - (\Sigma Y)^2/n = 79777 - (893)^2/10 = 32.1$
$L_{xy} = \Sigma XY - (\Sigma X)(\Sigma Y)/n = 71388-(799)(893)/10 = 37.3$

$V_{total} = V_{regression} + V_{residual}$ => $V_{total} = n-1$; $V_{regression} = 1$; $V_{residual} = n-2$;

2nd Calculate the r-value, do the following:

$$r = \frac{Lxy}{\sqrt{Lxx\ Lyy}} = \frac{37.3}{\sqrt{60.9 \times 32.1}} = 0.844$$

3rd Make the test for the correlation coefficient

Hypothesis of correlation coefficient

The r-value calculated is from the samples and is used for the estimation of the population. As there are the possibilities of sampling error, we should conduct a statistic test for the r and make the statistic significance judgments between the variable X and Y.

a. The hypothesis:

H_0: $\beta = 0$ (No linear correlation between the X and Y)
H_1: $\beta > 0$
$\alpha = 0.05$

b. The t-test and its formula are following:

$$t = \frac{r - 0}{Sr} = \frac{r}{\sqrt{(1 - r^2)/(n-2)}} = \frac{0.844}{\sqrt{(1-0.844^2)/(10-2)}} = 4.451$$

Where, Sr is standard error of r; n is number of samples.
Note: the population error is supposed $R_0=0$.

c. Find the p-value

From the t-table, with $V = 2$, we find $0.0025>p>0.001$; with $\alpha = 0.05$ and $\alpha > p$, we reject the $\boldsymbol{H_0}$ and have statistic significant confidence, with 95%, accepting the X-variable positive correlation with Y-variable in the linear regression mode.
(Note: as it is either positive or negative correlation, the single side of p-value applied)

Difference Between the Linear Equation and Linear Correlation

I type regression: Y variance is normal distribution and X variance has exact value.

II type regression: both of Y and X variance are normal distribution. It could be set up two regression equations:

$Y = a_{yx} + b_{yx} X$

$X = a_{xy} + b_{xy} Y$

Regression equation indicates the two of variance is co-existing. It is quantity measurable.

Regression correlation indicates the two of variance with possible relationship of correlation. It is correlation between the two set of variance in data.

Similarity Between the Linear Equation and Linear Correlation

r and b always have same positive or negative value.
r and b have the hypothesis test on same level.
r and b could be transferred each other.

I Type Regression:
(r => b)

$$r = \frac{\Sigma(X - X_m)\,\Sigma(Y - Y_m)}{\sqrt{\Sigma(X - X_m)^2\,\Sigma(Y - Y_m)^2}} = \frac{Lxy}{\sqrt{Lxx\,Lyy}} = b\sqrt{\frac{Lxx}{Lyy}}$$

II Type Regression:
(b ⇔ r)

$$b_{yx} = r\,(S_y / S_x)\;;\quad b_{xy} = r\,(S_x / S_y)\;;\quad r = \sqrt{b_{yx}\,b_{xy}}$$

where, S_y is the standard error of Y variance; S_x is the standard error of X variance;

Coefficient of Determination

r^2 is the Coefficient of Determination.

$r^2 = SS_{regression} / SS_{total}$
$SS_{regression} = r^2\,SS_{total}$

e.g. the value of r^2 could simply indicate how good of fit between the two variance. If r=0.20, the r^2 =0.04 and indicated $SS_{regression}$ only 4% weight power on SS_{total}.

Rank Data in Correlation

If the variance was not in a normal distribution or don't know the distribution or in a rank data, the rank correlation should be considered.

Spearman Method

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

where, d is the difference between the variance rank, n is number of "d" pairs. r_s is the estimation coefficient for ρ_s.

Example 5: Correlation coefficient of determination for rank data

Suppose we have a set of data as follows:

n	X	X Rank	Y	Y Rank	$d=X_R-Y_R$	d^2
1	0.7	1	21.5	3	-2	4
2	1.0	2	18.9	2	0	0
3	1.7	3	14.4	1	2	4
4	3.7	4	46.5	7	-3	9
5	4.0	5	27.3	4	1	1
6	5.1	6	64.6	9	-3	9
7	5.5	7	46.3	6	1	1
8	5.7	8	34.2	5	3	9
9	5.9	9	77.6	10	-1	1
10	10.0	10	55.1	8	2	4
					sum	42

1st Making the null hypothesis and setting the test confidence level:

H_0: $P_S = 0$, (No linear regression)
H_1: $P_S \neq 0$,
$\alpha = 0.05$

2nd Rank the X and Y as showed in the table above, the calculate the r value according to the following formula:

$$r_S = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6x42}{10(10^2-1)} = 0.745$$

3rd Find the p value

From the *Rs* -table, with $n = 3$, we find $p<0.02$; with $\alpha = 0.05$ and $\alpha > p$, we reject the H_0 and have statistic significant confidence, with 95%, accepting the X-variable positive correlated with Y-variable.

Correction for Approximation Error

Correction for continuity

If there are many of the same rank (tie rank), the adjustment formula for r's is needed as follows.

$$r'_s = \frac{[(n^3 - n)/6] - (Tx + Ty) - \Sigma d^2}{\sqrt{[(n^3 - n)/6] - 2Tx}\ \sqrt{[(n^3 - n)/6] - 2Ty}}$$

$Ty = \Sigma(t_y^3 - t_y)/n_y$

$Tx = \Sigma(t_x^3 - t_x)/n_x$

where, t_x and t_y are the number of the tie rank of variable X and Y respectively.

Example 6: calculation of correction for approximation error

Suppose the data in example 5 changed as follows:

1. No. 1 to No.5 with the tie rank, then the average rank would be: (1+2+3+4+5)/5 =3; and t=5.

2. No. 6 to No.8 with the tie rank, then the average rank would be: (6+7+8)/3 =7; and t=3.

3. No. 9 to No.10 with the tie rank, then the average rank would be: (9+10)/2 =9.5; and t=2.

   $n_x = 12$ (Total sample of X remains the same)
   Ty = 0 (No tie rank for variable Y)
   $\Sigma d^2 = 33.5$ (calculations was supposed)

Therefore,

$Tx = \Sigma(t_x^3 - t_x)/n_x = [(5^3-5)+(3^3-3)+(2^3-2)] / 12 = 12.5$

$$r'_s = \frac{[(n^3 - n)/6] - (Tx + Ty) - \Sigma d^2}{\sqrt{}\ \sqrt{}}$$

$$\sqrt{[(n^3 - n)/6] - 2T_x} \quad \sqrt{[(n^3 - n)/6] - 2T_y}$$

$$= \frac{[(10^3 - 10)/6] - (12.5 + 0) - 33.5}{\sqrt{[(10^3 - 10)/6] - 2(12.5)}\sqrt{[(10^3 - 10)/6] - 2(0)}} = 0.783$$

Find the p value

From the R_S -table, with $n = 10$, we find $0.02>p>0.01$; with $\alpha = 0.05$ and $\alpha > p$, we reject the $\boldsymbol{H_0}$ and have statistic significant confidence, with 95%, accepting the X-variable positive correlated with Y-variable.

Note: if we didn't calculate the correction for approximation error, the r-value would be:

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6x33.5}{10(10^2-1)} = 0.797$$

From the R_S -table, with $n = 10$, we find $0.01>p>0.005$; it would be quite difference with the r_s calculated by the correction for approximation error.

Other cases to apply the linear regression in understanding the effect

Effect of fertilizer on plant growths:
In a study measuring the influence of different quantities of fertilizer on plant growth, the independent variable would be the amount of fertilizer used. The dependent variable would be the growth in height or mass of the plant. The controlled variables would be the type of plant, the type of fertilizer, the amount of sunlight the plant gets, the size of the pots, etc.

Effect of drug dosage on symptom severity:
In a study of how different doses of a drug affect the severity of symptoms, a researcher could compare the frequency and intensity of symptoms when different doses are administered. Here the independent variable is the dose and the dependent variable is the frequency/intensity of symptoms.

Effect of temperature on pigmentation:
In measuring the amount of color removed from beetroot samples at different temperatures, temperature is the independent variable and amount of pigment removed is the dependent variable.

Effect of sugar added in a coffee:
The taste varies with the amount of sugar added in the coffee. Here, the sugar is the independent variable, while the taste is the dependent variable.

Other remarks**:**

The variables with no relationship should not use the regression equation or regression correlation.

Before the analysis of regression, it may be draw a draft scatter point to test for understand the data.

Confidence intervals
The formulas given in the previous section allow one to calculate the point estimates of "*a"* and "*b"* -- that is, the coefficients of the regression line for the given set of data. However, those formulas don't tell us how precise the estimates are, i.e., how much the estimators "*a"* and "*b*" vary from sample to sample for the specified sample size. Confidence intervals were devised to give a plausible set of values to the estimates one might have if one repeated the experiment a very large number of times.

The standard method of constructing confidence intervals for linear regression coefficients relies on the normality assumption, which is justified if either:

1. the errors in the regression are normally distributed (the so-called classic regression assumption), or
2. the number of observations ***n*** is sufficiently large, in which case the estimator is approximately normally distributed, or the case is justified by the central limit theorem.

Mathematically, we can use the variable transformation to be rectification based on the data category, e.g. curve fitting with variance transformation:

$Y=AB^x$ (B>0) or lgY=a+bX. These methods could extend the scope of application in linear regressions and please refer the other relevant readings.

Chapter 10

Multiply Linear Regression (MLR)

What Is Multiple Linear Regression (MLR)?

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

KEY TAKEAWAYS

1. Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

2. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.
3. MLR is used extensively in econometrics and financial inference.

What Multiple Linear Regression (MLR) Can Tell You

Simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables - an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions:

1. There is a linear relationship between the dependent variables and the independent variables.
2. The independent variables are not too highly correlated with each other. *"Yi"* observations are selected independently and randomly from the population.
3. Residuals should be normally distributed with a mean of *0* and variance σ.

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R^2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output

from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

How to Use Multiple Linear Regression (MLR)

As an example, an analyst may want to know how the movement of the market affects the price of ExxonMobil (XOM). In this case, their linear equation will have the value of the S&P 500 index as the independent variable, or predictor, and the price of XOM as the dependent variable.

In reality, there are multiple factors that predict the outcome of an event. The price movement of ExxonMobil, for example, depends on more than just the performance of the overall market. Other predictors such as the price of oil, interest rates, and the price movement of oil futures can affect the price of XOM and stock prices of other oil companies. To understand a relationship in which more than two variables are present, multiple linear regression is used.

Multiple linear regression (MLR) is used to determine a mathematical relationship among a number of random variables. In other terms, MLR examines how multiple independent variables are related to one dependent variable. Once each of the independent factors has been determined to predict the dependent variable, the information on the multiple variables can be used to create an accurate prediction on the level of effect they have on the outcome variable. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points.

Referring to the MLR equation above, in our example:

y_i = dependent variable - the price of XOM
x_{i1} = interest rates
x_{i2} = oil price
x_{i3} = value of S&P 500 index
x_{i4}= price of oil futures
B_0 = y-intercept at time zero
B_1 = regression coefficient that measures a unit change in the dependent variable when xi1 changes - the change in XOM price when interest rates change

B2 = coefficient value that measures a unit change in the dependent variable when xi2 changes - the change in XOM price when oil prices change

The least-squares estimates, B0, B1, B2, … Bp, are usually computed by statistical software. As many variables can be included in the regression model in which each independent variable is differentiated with a number - 1,2, 3, 4...p. The multiple regression model allows an analyst to predict an outcome based on information provided on multiple explanatory variables.

Still, the model is not always perfectly accurate as each data point can differ slightly from the outcome predicted by the model. The residual value, E, which is the difference between the actual outcome and the predicted outcome, is included in the model to account for such slight variations.

The Difference Between Linear and Multiple Regression

Ordinary linear squares (OLS) regression compares the response of a dependent variable given a change in some explanatory variables. However, it is rare that a dependent variable is explained by only one variable. In this case, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. Multiple regressions can be linear and nonlinear.

Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables.

Dependent and independent variables are variables in mathematical modeling, statistical modeling and experimental sciences. Dependent variables receive this name because, in an experiment, their values are studied under the supposition or hypothesis that they depend, by some law or rule (e.g., by a mathematical function), on the values of other variables. Independent variables, in turn, are not seen as depending on any other variable in the scope of the experiment in question; thus, even if the existing dependency is invertible (e.g., by finding the inverse function when it exists), the nomenclature is kept if the inverse dependency is not the object of study in the experiment. In this sense, some common independent variables are time, space, density, mass, fluid flow rate, and previous values

of some observed value of interest (e.g. human population size) to predict future values (the dependent variable).

Of the two, it is always the dependent variable whose variation is being studied, by altering inputs, also known as regressors in a statistical context. In an experiment, any variable that the experimenter manipulates can be called an independent variable. Models and experiments test the effects that the independent variables have on the dependent variables. Sometimes, even if their influence is not of direct interest, independent variables may be included for other reasons, such as to account for their potential confounding effect.

In an experiment, the variable manipulated by an experimenter is called an independent variable. The dependent variable is the event expected to change when the independent variable is manipulated.

In data mining tools (for multivariate statistics and machine learning), the dependent variable is assigned a role as target variable (or in some tools as label attribute), while an independent variable may be assigned a role as regular variable. Known values for the target variable are provided for the training data set and test data set, but should be predicted for other data. The target variable is used in supervised learning algorithms but not in unsupervised learning.

Multiple Linear Regression Equation

In mathematical modeling, the dependent variable is studied to see if and how much it varies as the independent variables vary. In the multiple linear model:

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Where, $\hat{Y}$ (dependent variable) is the estimated value for Y-variance; X_1, X_2, … X_n are the X-variance (independent variable); a is a coefficient (constant); b_1, b_2,…b_n are partial regression coefficient.

e.g. in a local area, the 10-year-old children have showed their Vital Capacity (Y) with their Height (cm), (X_1), and Weight (kg), (X_2), as the following multiple regression equation:

$Y^ = - 0.5657 + 0.0050X_1 + 0.0541X_2$

Where,
a = - 0.5657; b_1 = 0.0050, it indicates that the Height (X_1) increase 1cm and the Vital Capacity (Y^) would have 0.0050 increases of measurable effect in the groups of children, as the Weight (X_2) remains in the same or excludes the effects of Weight (X_2) factor.

Multiple regression could measure or forecast the one factor (independent variable) on how much effects with another factors (dependent variable) respectively.

Methods to set up multiple linear regression equation

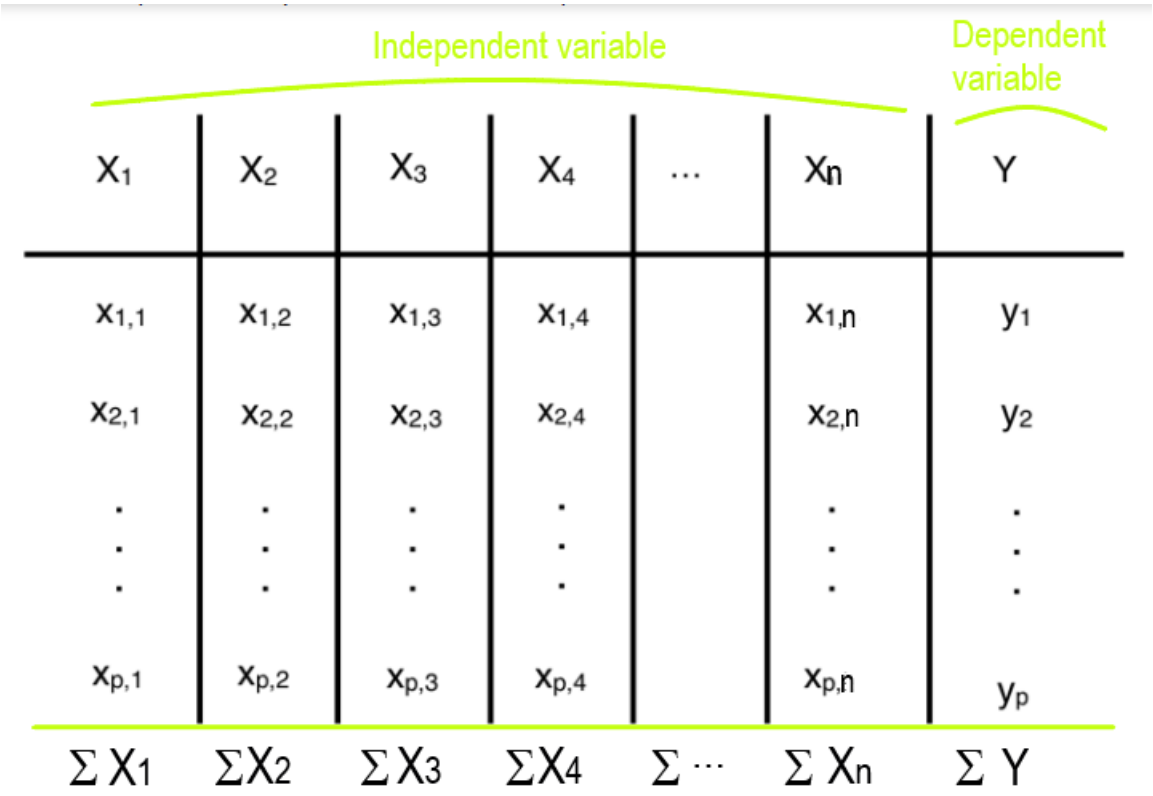
A major application of matrices is to represent linear transformations, that is,

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2 \\ \vdots \qquad\qquad \vdots \\ a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n = b_m \end{cases}$$

In mathematics, a matrix (plural matrices) is a rectangular array or table (see irregular matrix) of numbers, symbols, or expressions, arranged in rows and columns. Matrices are commonly written in box brackets or parentheses:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

The calculation is illustrated as follows:



Calculate the ΣX_1, ΣX_2, ΣX_3, … ΣX_n, ΣY.
Calculate the ΣX^2_1, ΣX^2_2, ΣX^2_3, … ΣX^2_n, ΣY^2.
Calculate the $\Sigma X_1 Y$, $\Sigma X_2 Y$, $\Sigma X_3 Y$, … $\Sigma X_n Y$.

Example 1: How to set up the multiple linear regression equation

$\hat{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$

Suppose we have a set of data: in a local area, the 10-year-old children have showed their Vital Capacity (Y) with their Height (cm), (X_1), and Weight (kg), (X_2), as the follows:

	X_1 (cm)	X_1^2	X_2 (kg)	X_2^2	Y (L)	Y^2	$X_1 * X_2$	$X_1 Y$	$X_2 Y$
1	135.1	18252.0	32.0	1024.0	1.75	3.06	4323.2	236.4	56.0
2	139.9	19572.0	30.4	924.2	2.00	4.00	4253.0	279.8	60.8
3	163.6	26765.0	46.2	2134.4	2.75	7.56	7558.3	449.9	127.1
4	146.5	21462.3	33.5	1122.3	2.50	6.25	4907.8	366.3	83.8
5	156.2	24398.4	37.1	1376.4	2.75	7.56	5795.0	429.6	102.0
6	156.4	24461.0	35.5	1260.3	2.00	4.00	5552.2	312.8	71.0
7	167.8	28156.8	41.5	1722.3	2.75	7.56	6963.7	461.5	114.1
8	149.7	22410.1	31.0	961.0	1.50	2.25	4640.7	224.6	46.5

9	145.0	21025.0	33.0	1089.0	2.50	6.25	4785.0	362.5	82.5
10	148.5	22052.3	37.2	1383.8	2.25	5.06	5524.2	334.1	83.7
11	165.5	27390.3	49.5	2450.3	3.00	9.00	8192.3	496.5	148.5
12	135.0	18225.0	27.6	761.8	1.25	1.56	3726.0	168.8	34.5
13	153.3	23500.9	41.0	1681.0	2.75	7.56	6285.3	421.6	112.8
14	152.0	23104.0	32.0	1024.0	1.75	3.06	4864.0	266.0	56.0
15	160.5	25760.3	47.2	2227.8	2.25	5.06	7575.6	361.1	106.2
16	153.0	23409.0	32.0	1024.0	1.75	3.06	4896.0	267.8	56.0
17	147.6	21785.8	40.5	1640.3	2.00	4.00	5977.8	295.2	81.0
18	157.5	24806.3	43.3	1874.9	2.25	5.06	6819.8	354.4	97.4
19	155.1	24056.0	44.7	1998.1	2.75	7.56	6933.0	426.5	122.9
20	160.5	25760.3	37.5	1406.3	2.00	4.00	6018.8	321.0	75.0
21	143.0	20449.0	31.5	992.3	1.75	3.06	4504.5	250.3	55.1
22	149.4	22320.4	33.9	1149.2	2.25	5.06	5064.7	336.2	76.3
23	160.8	25856.6	40.4	1632.2	2.75	7.56	6496.3	442.2	111.1
24	159.0	25281.0	38.5	1482.3	2.50	6.25	6121.5	397.5	96.3
25	158.2	25027.2	37.5	1406.3	2.00	4.00	5932.5	316.4	75.0
26	150.0	22500.0	36.0	1296.0	1.75	3.06	5400.0	262.5	63.0
27	144.5	20880.3	34.7	1204.1	2.25	5.06	5014.2	325.1	78.1
28	154.6	23901.2	39.5	1560.3	2.50	6.25	6106.7	386.5	98.8
29	156.5	24492.3	32.0	1024.0	1.75	3.06	5008.0	273.9	56.0
(n=29)									
Total: Σ	4424.7	677060.4	1076.7	40832.4	64.00	146.88	165239.8	9826.7	2427.3
Mean	152.6		37.1		2.21				

1st Calculations and summaries:

$\Sigma X_1 = 4427.7$ $\Sigma X_2 = 1076.7$ $\Sigma Y = 64.$

$\Sigma X^2_1 = 677060.4$ $\Sigma X^2_2 = 40832.4$ $\Sigma Y^2 = 146.88$

$\Sigma X_1X_2 = 165239.8$ $\Sigma X_1Y = 9826.7$ $\Sigma X_2Y = 2427.3$

2nd Use the following matrix methods to calculate:

$$A = \begin{pmatrix} n & \Sigma X_1 & \Sigma X_2 & \cdots & \Sigma X_m \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1X_2 & \cdots & \Sigma X_1X_m \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \Sigma X_m & \Sigma X_mX_1 & \Sigma X_mX_2 & \cdots & \Sigma X_m^2 \end{pmatrix} = XX, \qquad B = \begin{pmatrix} \Sigma Y \\ \Sigma X_1Y \\ \vdots \\ \Sigma X_mY \end{pmatrix} = X^tY.$$

For the case of two impendent variances, its matrix would be as follows:

$$A = \begin{bmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X^2_1 & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X^2_2 \end{bmatrix} \qquad B = \begin{bmatrix} \Sigma Y \\ \Sigma X_1 Y \\ \Sigma X_2 Y \end{bmatrix}$$

Place the number to the matrix as follows:

$$A = \begin{bmatrix} 29 & 4427.7 & 1076.7 \\ 4424.7 & 677060.4 & 165239.8 \\ 1076.7 & 165239.8 & 40832.4 \end{bmatrix} \qquad B = \begin{bmatrix} 64 \\ 9826.7 \\ 2427.3 \end{bmatrix}$$

After several calculations*, the inverse matrix of A would be as follows:

$$1 / A = \begin{bmatrix} 15.632356 & -0.126109 & 0.098131 \\ -0.126109 & 0.001137 & -0.001275 \\ 0.098131 & -0.001275 & 0.002597 \end{bmatrix}$$

and

$$b = \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} = (1/A) \times B = \begin{bmatrix} -0.565664 \\ 0.005017 \\ 0.054061 \end{bmatrix}$$

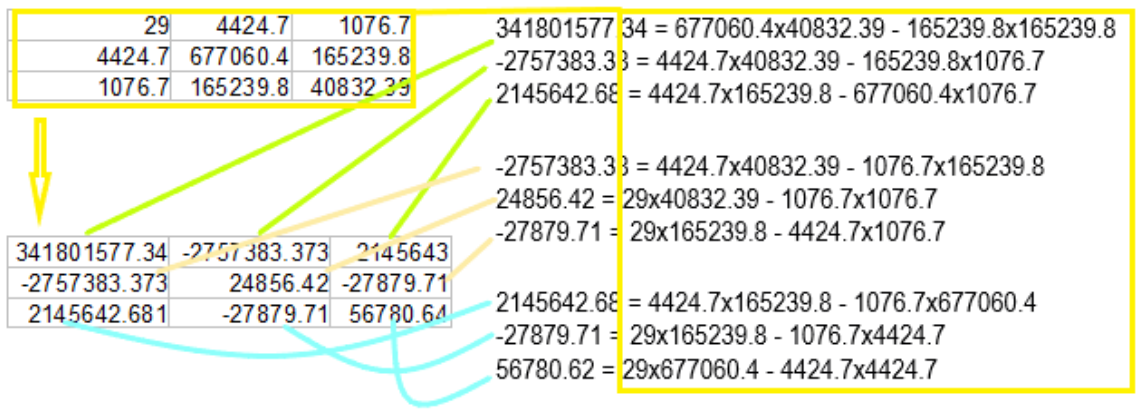
The multiple regression equation:

$$\hat{Y} = -0.5657 + 0.0050X_1 + 0.0541X_2$$

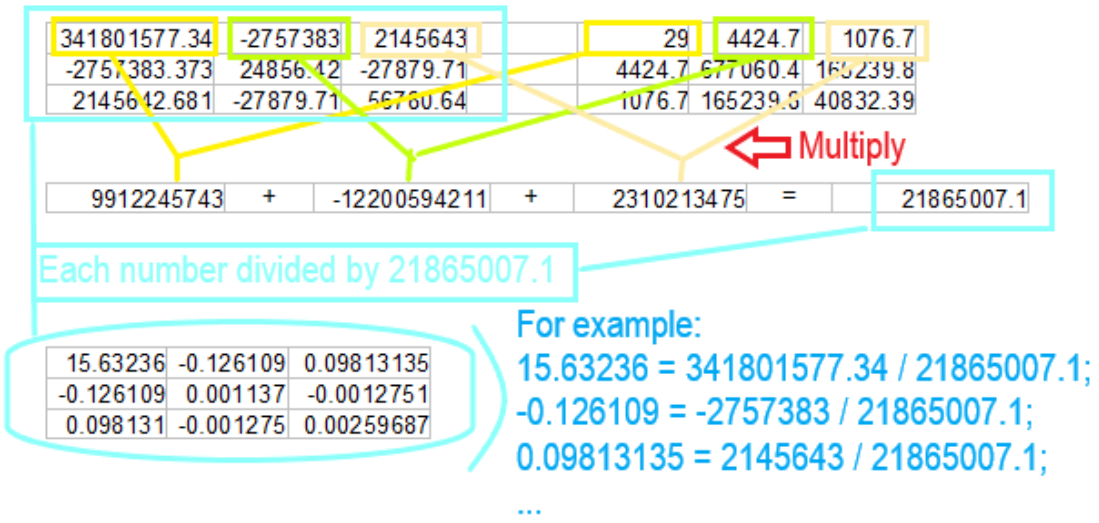
*Note: the following figure showed a simple method on how to make a change between A and 1/A.

In this case, the calculation in details is illustrated as follows:

1st



2nd & 3rd



The steps of calculation:

1st, Calculate the number or value in each cell of the table by matrix methods, as it illustrated in the number from yellow table to the number in light grey table;

2nd, Multiple the number in the light grey table by the number in yellow table, one by one, in the first row by the order respectively;

3rd, Add the three numbers in the small light grey table to be: 21865007.1;

Finally, Use the number in the light green table divided by the number of 21865007.1, one by one, and fill out each cell of the table with the calculating results by the order respectively; The numbers in the new light green table are the inverse matrix of A (or 1/A).

The other methods for the multiple linear regression equation:

$$\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} & \cdots & x_{1,N} \\ 1 & x_{2,1} & x_{2,2} & x_{2,3} & \cdots & x_{2,N} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{p,1} & x_{p,2} & x_{p,3} & \cdots & x_{p,N} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_N \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_p \end{pmatrix}$$

The large letters are the matrices and the smaller letters describe the dimensions of each term. We are solving for the beta vector. After some transformations, this can be expressed as:

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

? is a vector and from this vector we can take our required values to construct the desired equation.

$$\boldsymbol{\beta_0} + \boldsymbol{\beta_1} \boldsymbol{X_1} + \boldsymbol{\beta_2} \boldsymbol{X_2} + \cdots + \boldsymbol{\beta_N} \boldsymbol{X_N} = \boldsymbol{Y}$$

This equation can then be used to make predictions on data where the values of **Y** are unknown.

Multiple Linear Regression Analysis

Multiple regression analysis is based on the multiple linear regression equation. The linear hypnosis test is testing the u (error term).

Multiple Regression Equation:

$Y^{\wedge} = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n$

Multiple Regression Analysis

- Multiple Regression:
- $Y^{\wedge} = a + b_1X_1 + b_2X_2 + B_3X_3 + \ldots + B_nX_n + u$

Where:
Y= the variable that we are trying to predict(DV)
X= the variable that we are using to predict Y(IV)
a= the intercept
b= the slope (Coefficient of X1)
u= the regression residual (error term)

$$\Sigma (Y - Y_m)^2 = \Sigma (Y^{\wedge} - Y_m)^2 + \Sigma (Y - Y^{\wedge})^2$$

$$SS_{total} = SS_{regression} + SS_{residual}$$

$$SS_{total} = L_{yy} = \Sigma(Y - Y_m)^2 = \Sigma Y^2 - (Y)^2/n$$

$$SS_{regression} = \Sigma (Y^{\wedge} - Y_m)^2 = b' B - (Y)^2/n$$

$$SS_{residual} = SS_{total} - SS_{regression}$$

Example 2: Methods of Analysis of Variance (ANOV)

Suppose we have the data list in example 1 above.

1st Making the null hypothesis and setting the test confidence level:

H_0: $b_1 = 0, b_2 = 0, \ldots b_n = 0$ *(No multiple linear regression)*
H_1: $b_1 \neq 0, b_2 \neq 0, \ldots b_n \neq 0,$

α = 0.05

2nd From the case above, to calculate it, do the following:

$$F = \frac{SS_{regression} / V_{regression}}{SS_{residual} / (n-m-1)}$$

(Note: the formula is the similar with F-test in thesimple linear regression)

$$SS_{total} = L_{yy} = \Sigma(Y - Y_m)^2 = \Sigma Y^2 - (Y)^2/n = 146.88 - (64)^2/29 = 5.6336$$

$$SS_{regression} = \Sigma (Y^{\wedge} - Y_m)^2 = b' B - (Y)^2/n$$

$$= (-0.565664 \quad 0.005017 \quad 0.054061) \begin{bmatrix} 64 \\ 9826.65 \\ 2427.325 \end{bmatrix} - (64)^2/29 = 3.0800$$

$$SS_{residual} = SS_{total} - SS_{regression} = 5.6336 - 3.0800 = 2.5536$$

$$V_{regression} = 2;\ V_{residual} = (29 - 2 - 1) = 26$$

$$F = \frac{SS_{regression} / V_{regression}}{SS_{residual} / (n-m-1)} = \frac{3.0800 / 2}{2.5536 / 26} = 15.680$$

3rd Find the p value

From the *F* -table, with v = 26 and v =2, we find 0.01>p>; with α = 0.05 and α >p, we reject the $\boldsymbol{H_0}$ and have statistic significant confidence, with 95%, accepting the effects of X1-variable and X2-variable correlated with Y-variable in the multiple linear regression.

Coefficient of Partial Regression

If the F-test above indicated that the multiple regression is confidential existed, the further test for partial regression coefficient could be conducted.

The purpose of the test is determining how a factor to be relevant.

Example 3: Calculation and hypothesis of partial regression coefficient

Suppose we have the data list in example 1 above.

1st Making the null hypothesis and setting the test confidence level:

The hypnosis test for b1, b2 (t-test)

H_0: $b_1 = 0, b_2 = 0$
H_1: $b_1 \neq 0, b_2 \neq 0$
$\alpha = 0.05$

2nd To calculate it, do the following:

$$t_i = \frac{b_i - b_t}{\sqrt{SS_{residual} / (n-m-1)} \times \sqrt{C_{ii}}}$$

where, Cii is the i-row and i-column of matrix A.

$SS_{residual} = 2.5536$. $C_{11} = 0.001137$, $b_t = 0$ (H_0)

$$t_1 = \frac{b_1}{\sqrt{SS_{residual} / (n-m-1)} \times \sqrt{C_{ii}}} = \frac{0.0050}{\sqrt{2.5536 / (29-2-1)} \times \sqrt{0.001137}}$$

$= 0.473$

$$t_2 = \frac{b_2}{\sqrt{SS_{residual} / (n-m-1)} \times \sqrt{C_{ii}}} = \frac{0.0541}{\sqrt{2.5536 / (29-2-1)} \times \sqrt{0.002597}}$$

$= 3.387$

$V_{residual}$ = 26, t-table, t_1: p>0.05; t_2: 0.005>p>0.002. α = 0.05, not refuse the $H_{0,1}$ of b_1=0; and refuse the $H_{0,2}$ and accept $H_{1,2}$.

3rd Find the p value

From the *t* -table, with *v* = 26, we find t_1: p_1>0.05; t_2: 0.005>p_2>0.002; with α = 0.05 and α >p_1, we do not reject the $H_{0(b1)}$ and reject $H_{0(b2)}$. It means that we have statistic significant confidence, with 95%, accepting the no-effects of X_1-variable correlated with Y-variable; and the effects of X_2-variable correlated with Y-variable in multiple linear regression.

Example 4: Calculation and test the simple linear regression coefficient

Further, we can exclude the height variance (X_1) and use simple linear regression to test the correlation of the weight variance (X_2) with variable (Y).

The methods applied are the similar as illustrated in the simple linear regression.

Sum up the calculation as follows:

n=29, ΣX_2 = 1076.7, ΣX^2_2 = 40832.4, ΣY= 64, $\Sigma X_2 Y$ = 2427.3, ΣY^2 = 146.88.

$L_{xx} = L_{22} = \Sigma X^2_2 - (\Sigma X_2)^2/n$ = 40832.4 – $(1076.7)^2/29$ = 857.1179

$L_{xy} = L_{2y} = \Sigma X_2 Y - (\Sigma X_2 \Sigma Y)/n$ = 2427.325 – (1076.7)(64)/29 = 51.1595

$X_m = X_{2m}$ = 1076.7/29 =37.1276

Y_m = 64/29 =2.2069

$b = L_{xy} / L_{xx}$ = 51.1595/857.1179 = 0.0597

$a = Y_m - bX_m$ = 2.2069 – (0.0597)(37.1276) = -0.0096

Therefore, the equation is: $Y^ = -0.0096 + 0.0597X_2$

To Test the b coefficient

1st Making the null hypothesis and setting the test confidence level:

H_0: $b = 0$
H_1: $b \neq 0$
$\alpha = 0.05$

2nd To calculate it, do the following:

$SS_{total} = L_{yy} = \Sigma(Y - Y_m)^2 = \Sigma Y^2 - (Y)^2/n = 146.88 - (64)^2/29 = 5.6336$

$SS_{regression} = L_{xy} / L_{xx} = (51.1595)^2 / 857.1179 = 3.0536$

$SS_{residual} = SS_{total} - SS_{regression} = 5.6336 - 3.0536 = 2.5800$

$V_{residual} = n - 2 = 29 - 2 = 27$

$$S_{yx} = \sqrt{SS_{residual} / n} = \sqrt{2.5800 / 27} = 0.3091$$

$$t = \frac{b}{S_b} = \frac{b}{S_{yx} / \sqrt{L_{xx}}} = \frac{0.0597}{0.3091 / \sqrt{857.1179}} = 5.655$$

3rd Find the p value

From the *t* -table, with $v = 27$, we find $0.001 > p$; with $\alpha = 0.05$ and $\alpha > p$, we reject H_0. It means that we have statistic significant confidence, with 95%, accepting the effects of X_2-variable correlated with Y-variable in the linear regression.

Correlation of Multiple Linear

Coefficient of Multiple Correlation (R)

In statistics, the coefficient of multiple correlation is a measure of how well a given variable can be predicted using a linear function of a set of other variables. It is the correlation between the variable's values and the best predictions that can be computed linearly from the predictive variables.

The coefficient of multiple correlation takes values between .00 and 1.00; a higher value indicates a high predictability of the dependent variable from the independent variables, with a value of 1 indicating that the predictions are exactly correct and a value of 0 indicating that no linear combination of the independent variables is a better predictor than is the fixed mean of the dependent variable.

The coefficient of multiple correlation is known as the square root of the coefficient of determination, but under the particular assumptions that an intercept is included and that the best possible linear predictors are used, whereas the coefficient of determination is defined for more general cases, including those of nonlinear prediction and those in which the predicted values have not been derived from a model-fitting procedure.

The coefficient of multiple correlation, denoted R, is a scalar that is defined as the Pearson correlation coefficient between the predicted and the actual values of the dependent variable in a linear regression model that includes an intercept.

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{residual}}{SS_{total}}$$

$$R = \sqrt{\frac{SS_{regression}}{SS_{total}}} = \sqrt{1 - \frac{SS_{residual}}{SS_{total}}}$$

Example 5: Test of coefficient of multiple correlation

In case above, we calculate it as follows:

$$SS_{regression} = \Sigma (Y\hat{} - Y_m)^2 = b' B - (Y)^2/n$$

$$= (-0.565664 \quad 0.005017 \quad 0.054061) \begin{bmatrix} 64 \\ 9826.65 \\ 2427.325 \end{bmatrix} - (64)^2/29 = 3.0800$$

$SS_{total} = L_{yy} = \Sigma(Y - Y_m)^2 = \Sigma Y^2 - (Y)^2/n = 146.88 - (64)^2/29 = 5.6336$

$$R = \sqrt{\frac{SS_{regression}}{SS_{total}}} = \sqrt{\frac{3.0800}{5.6336}} = 0.7394$$

Test of Coefficient of Multiple Correlations

1st Making the null hypothesis and setting the test confidence level:

H_0: $C = 0$ (Coefficient of multiple correlation =0)
H_1: $C \neq 0$
$\alpha = 0.05$

2nd To calculate it, do the following:

F-test

$$F = \frac{R^2}{1 - R^2} \times \frac{(n-m-1)}{m} = \frac{(0.7394)^2}{1 - (0.7394)^2} \times \frac{29-2-1}{2} = 15.679$$

3rd Find the p value

From the *F* -table, with $v = 2$, we find $0.01 > p$; with $\alpha = 0.05$ and $\alpha > p$, we reject H_0. It means that we have statistic significant confidence, with 95%, accepting the correlation effects in the linear regression.

Coefficient of Partial Correlation

In probability theory and statistics, partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed. If we are interested in finding to

what extent there is a numerical relationship between two variables of interest, using their correlation coefficient will give misleading results if there is another, confounding, variable that is numerically related to both variables of interest. This misleading information can be avoided by controlling for the confounding variable, which is done by computing the partial correlation coefficient.

In probability theory and statistics, partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed.

If we are interested in finding to what extent there is a numerical relationship between two variables of interest, using their correlation coefficient will give misleading results if there is another, confounding, variable that is numerically related to both variables of interest. This misleading information can be avoided by controlling for the confounding variable, which is done by computing the partial correlation coefficient. This is precisely the motivation for including other right-side variables in a multiple regression; but while multiple regression gives unbiased results for the effect size, it does not give a numerical value of a measure of the strength of the relationship between the two variables of interest.

For example, if we have economic data on the consumption, income, and wealth of various individuals and we wish to see if there is a relationship between consumption and income, failing to control for wealth when computing a correlation coefficient between consumption and income would give a misleading result, since income might be numerically related to wealth which in turn might be numerically related to consumption; a measured correlation between consumption and income might actually be contaminated by these other correlations. The use of a partial correlation avoids this problem.

Like the correlation coefficient, the partial correlation coefficient takes on a value in the range from -1 to 1. The value -1 conveys a perfect negative correlation controlling for some variables (that is, an exact linear relationship in which higher values of one variable are associated with lower values of the other); the value 1 conveys a perfect positive linear relationship, and the value 0 conveys that there is no linear relationship.

The partial correlation coincides with the conditional correlation if the random variables are jointly distributed as the multivariate normal, other elliptical, multivariate hypergeometric, multivariate negative hypergeometric, multinomial or Dirichlet distribution, but not in general otherwise.

Semipartial Correlation (Part Correlation)

The semipartial (or part) correlation statistic is similar to the partial correlation statistic. Both compare variations of two variables after certain factors are controlled for, but to calculate the semipartial correlation one holds the third variable constant for either X or Y but not both, whereas for the partial correlation one holds the third variable constant for both. The semipartial correlation compares the unique variation of one variable (having removed variation associated with the Z variable(s)), with the unfiltered variation of the other, while the partial correlation compares the unique variation of one variable to the unique variation of the other.

The semipartial (or part) correlation can be viewed as more practically relevant "because it is scaled to (i.e., relative to) the total variability in the dependent (response) variable." Conversely, it is less theoretically useful because it is less precise about the role of the unique contribution of the independent variable.

The absolute value of the semipartial correlation of X with Y is always less than or equal to that of the partial correlation of X with Y. The reason is this: Suppose the correlation of X with Z has been removed from X, giving the residual vector ex . In computing the semipartial correlation, Y still contains both unique variance and variance due to its association with Z. But ex , being uncorrelated with Z, can only explain some of the unique part of the variance of Y and not the part related to Z. In contrast, with the partial correlation, only ey (the part of the variance of Y that is unrelated to Z) is to be explained, so there is less variance of the type that *ex* cannot explain.

Example 6: Calculation of Coefficient of Partial Correlation

In case above, we calculate it as follows:

n = 29

$\Sigma X_1 = 4427.7$ $\Sigma X_2 = 1076.7$ $\Sigma Y = 64.$

$\Sigma X^2_1 = 677060.4$ $\Sigma X^2_2 = 40832.4$ $\Sigma Y^2 = 146.88$

$\Sigma X_1X_2 = 165239.8$ $\Sigma X_1Y = 9826.7$ $\Sigma X_2Y = 2427.3$

$L_{11} = \Sigma X^2_1 - (\Sigma X_1)^2 / n = 677060.4 - (4427.7)^2 / 29 = 1957.9531$

$L_{22} = \Sigma X^2_2 - (\Sigma X_2)^2 / n = 40832.4 - (1076.7)^2 / 29 = 857.1179$

$L_{yy} = \Sigma Y^2 - (Y)^2/n = 146.88 - (64)^2 / 29 = 5.6336$ (=SS total)

$L_{12} = \Sigma X_1X_2 - (\Sigma X_1 \Sigma X_2) / n = 165239.8 - (4427.7)(1076.7) / 29 = 961.3693$

$L_{y1} = \Sigma X_1Y - (\Sigma X_1 \Sigma Y)/n = 9826.7 - (64)(4427.7) / 29 = 61.7948$

$L_{y2} = \Sigma X_2Y - (\Sigma X_2 \Sigma Y)/n = 2427.3 - (64)(1076.7) / 29 = 51.1595$

$r_{12} = L_{12} / \sqrt{L_{11} L_{22}} = 961.3693 / \sqrt{1957.9531 \times 857.1179} = 0.7421$

$r_{y1} = L_{y1} / \sqrt{L_{yy} L_{11}} = 61.7948 / \sqrt{5.6336 \times 1957.9531} = 0.5884$

$r_{y2} = L_{y2} / \sqrt{L_{yy} L_{22}} = 51.1595 / \sqrt{5.6336 \times 857.1179} = 0.7362$

$r_{12.y} = (r_{12} - r_{y1} r_{y2}) / \sqrt{(1 - r^2_{y1})(1 - r^2_{y2})}$

$= (0.7421 - 0.5884 \times 0.7362) / \sqrt{(1 - 0.5884^2)(1 - 0.7362^2)} = 0.5645$

$r_{y1.2} = (r_{y1} - r_{12} r_{y2}) / \sqrt{(1 - r^2_{12})(1 - r^2_{y2})}$

$= (0.5884 - 0.7421 \times 0.7362) / \sqrt{(1 - 0.7421^2)(1 - 0.7362^2)} = 0.0927$

$r_{y2.1} = (r_{y2} - r_{y1} r_{12}) / \sqrt{(1 - r^2_{y1})(1 - r^2_{12})}$

$= (0.7362 - 0.5884 \times 0.7421) / \sqrt{(1 - 0.5884^2)(1 - 0.7421^2)} = 0.5527$

Example 7: Test of Coefficient of Partial Correlation

1st Making the null hypothesis and setting the test confidence level:

H_0: $p_{12.y} = 0$, $p_{y2.1} = 0$,
H_1: $p_{12.y} \neq 0$, $p_{y2.1} \neq 0$,
$\alpha = 0.05$

2nd To calculate it, do the following:

n=29, m=2, v= 29 -2-1=26; $r_{y1.2} = 0.0927$, $r_{y2.1} = 0.5527$

$t_1 = r_{y1.2} / \sqrt{(1 - r^2_{y12})} \times \sqrt{n-m-1} = 0.0927 / \sqrt{(1-0.0927)^2} \times \sqrt{29-2-1} = 0.475$

$t_2 = r_{y2.1} / \sqrt{(1 - r^2_{y21})} \times \sqrt{n-m-1} = 0.5527 / \sqrt{(1-0.5527)^2} \times \sqrt{29-2-1} = 3.382$

3rd Find the p value

From the *t* -table, with $v = 26$, we find t_1: $p>0.50$; $\alpha = 0.05$, not reject the H_0 of $p_{12.y} = 0$; t_2: $0.002>p>0.001$; $\alpha = 0.05$, reject the H_0 of $p_{y2.1} = 0$.
It means that we have statistic significant confidence, with 95%, accepting the correlation effects of X2-variable with Y-variable in the linear regression when the X1-variable remains the same level.

The results of the coefficient correlation test is very similar with the coefficient regression test illustrated before, as they have the same set of data being tested; but one is for correlation and one is for regression.

The summary of the correlation and partial correlation was list on the following table:

Variance	Coefficient of Multiple Correlation	Coefficient of Partial Correlation
X1 (cm)	0.5884	0.0927
X2 (kg)	0.7362	0.5527

Though the Coefficient of Multiply Correlation indicated that there is the correlation among the variance, the Coefficient of Partial Correlation

indicated that the X2-indepent-variance has the linear regression with Y-dependent-variance, when X1-indepent-variance is fixed.

The examples are enabling you to compare the predicted values with the actual values. Not very surprisingly, the performance of the model on the training data looks quite convincing. In a real-world application, you would choose a setup with a hold-out set or cross-validation to determine the actual model performance. An actual one could be as an attractive model as the following graphs.