Statistics for Data Science

Dr. Harsh Dhiman

Department of AI & ML
Symbiosis Institute of Technology, Pune

May 2, 2023

॥वसुधैव कुटुम्बकम्॥
SYMBIOSIS INSTITUTE OF TECHNOLOGY (SIT)
Constituent of Symbiosis International (Deemed University), Pune
(Established under Section 3 of the UGC Act of 1956 vide notification number F-9-12/2001-U-3 of the Government of india)
Re-Accredited by NAAC with 'A' Grade

Outline

1. Evaluation Criteria
2. Chapter-I
Measures of Central Tendency & Dispersion
3. Chapter-II
Bivariate Analysis
4. Module-III
Sampling Techniques
5. Module-IV
Inferential Statistics: Hypothesis Testing

Evaluation Criteria

Evaluation Criteria for Internal Assessment

Component	Description	Date	Weightage
I	Viva	February	12.5%
II	Unit Test	March	25%
III	Industry Use-Case	March-April	37.5%
IV	Tutorial	Continuous	25%

Chapter-I
Measures of Central Tendency & Dispersion

Measures of Central Tendency & Dispersion

- Statistics is the branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of numerical data.
- It is divided into two parts:
  (a) Descriptive Statistics
  (b) Inferential Statistics

Machine Learning

- Machine learning is the branch of computer science that utilizes past experience to learn from and use its knowledge to make future decisions.
- Machine learning is at the intersection of computer science, engineering, and statistics.
- The goal of machine learning is to generalize a detectable pattern or to create an unknown rule from given examples

Statistical Modeling & Machine Learning Modeling

Statistical Modeling

- Formalization of relationships between variables in the form of mathematical equations.
- Required to assume shape of the model curve prior to perform model fitting on the data (for example, linear, polynomial, and so on).
- Statistical model predicts the output with accuracy of 85 percent and having 90 percent confidence about it.
- In statistical modeling, various diagnostics of parameters are performed, like p-value, and so on.

Contd...

Machine Learning Modeling

- Algorithm that can learn from the data without relying on rule-based programming
- Does not need to assume underlying shape, as machine learning algorithms can learn complex patterns automatically based on the provided data.
- Machine learning just predicts the output with accuracy of 85 percent.
- Machine learning models do not perform any statistical diagnostic significance tests

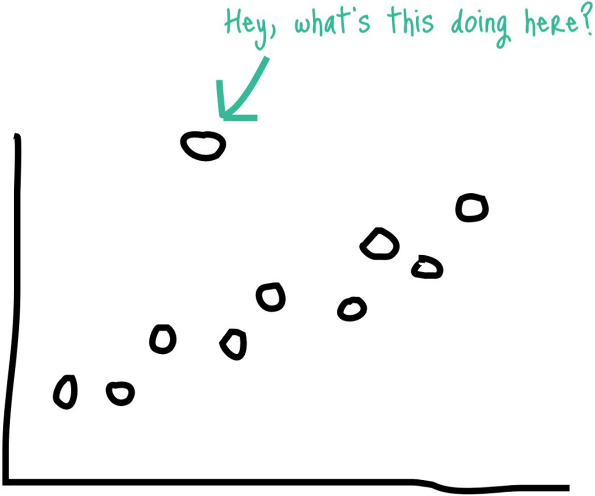
Keywords in Statistics

1. Population: This is the totality, the complete list of observations, or all the data points about the subject under study. People of Pune is an example of Population for a study that concerns Pune only.
2. Sample: A sample is a subset of a population, usually a small portion of the population that is being analyzed. For example, list of all the Punjabi's living in the city of Pune.
3. Parameter versus statistic: Any measure that is calculated on the population is a parameter, whereas on a sample it is called a statistic.

Don't be a MEAN

- Mean: This is a simple arithmetic average, which is computed by taking the aggregated sum of values divided by a count of those values.
- The mean is sensitive to outliers in the data.
- An outlier is the value of a set or column that is highly deviant from the many other values in the same data; it usually has very high or low values.

Outlier in Statistics



Contd...

1. Mathematically, sample mean can be expressed as

   $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{1}$$

   where n refers to the total number of elements present in a sample.

2. When total number of elements is written in CAPS(or uppercase), like N, it refers to a population mean.

Weighted Mean

1. This takes into account a weight for each element present in a given sample/population and can be given as

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

Median

1. The median is the middle number on a sorted list of the data.
2. If there is an even number of data values, the middle value is one that is not actually in the data set, but rather the average of the two values that divide the sorted data into upper and lower halves.
3. Compared to the mean, which uses all observations, the median depends only on the values in the center of the sorted data.

Contd...

1. Let's say we want to look at typical household incomes in neighborhoods around Lake Washington in Seattle.
2. In comparing the Medina neighborhood to the Windermere neighborhood, using the mean would produce very different results because Bill Gates lives in Medina.
3. If we use the median, it won't matter how rich Bill Gates is—the position of the middle observation will remain the same.

Outliers

1. The median is referred to as a robust estimate of location since it is not influenced by outliers (extreme cases) that could skew the results.
2. An outlier is any value that is very distant from the other values in a data set.
3. The exact definition of an outlier is somewhat subjective, although certain conventions are used in various data summaries and plots

Anomaly Detection

In contrast to typical data analysis, where outliers are sometimes informative and sometimes a nuisance, in anomaly detection the points of interest are the outliers, and the greater mass of data serves primarily to define the "normal" against which anomalies are measured.

Estimates of Variability

1. Location is just one dimension in summarizing a feature.
2. A second dimension, variability, also referred to as dispersion, measures whether the data values are tightly clustered or spread out.
3. At the heart of statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability, and making decisions in the presence of it.

Contd...

1. Deviations The difference between the observed values and the estimate of location.Synonyms errors, residuals
2. Variance The sum of squared deviations from the mean divided by $n - 1$ where n is the number of data values. Synonyms mean-squared-error
3. Standard deviation The square root of the variance. Synonyms l2-norm, Euclidean norm
4. Mean absolute deviation The mean of the absolute value of the deviations from the mean. Synonyms: l1-norm, Manhattan norm
5. Range The difference between the largest and the smallest value in a data set.
6. Percentile The value such that P percent of the values take on this value or less and (100–P) percent take on this value or more. Synonyms quantile

$\mathcal{L}_1$-norm

1. Consider a vector $x = [x_1, x_2, \ldots, x_n]$, the $\mathcal{L}_1$-norm is given as

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| = |x_1| + |x_2| + \ldots + |x_n| \tag{3}$$

2. A norm is a mathematical thing that is applied to a vector (like the vector **x** above).
3. The norm of a vector maps vector values to values in $[0, \infty)$.
4. In machine learning, norms are useful because they are used to express distances: this vector and this vector are so-and-so far apart, according to this-or-that norm.

$\mathcal{L}_2$-Norm

1. The simplest norm conceptually is Euclidean distance. This is what we typically think of as distance between two points in space:

$$\|x\|_2 = \sqrt{\left(\sum_i x_i^2\right)} = \sqrt{x_1^2 + x_2^2 + \ldots + x_i^2} \tag{4}$$

Comparison between $\mathcal{L}_1$ and $\mathcal{L}_2$ Norms

Aspect	$\mathcal{L}_1$	$\mathcal{L}_2$
Robustness	High	Low
Stability	Low	High
Time complexity	High	Low
Sparsity	High	Low

Standard Deviation & Variance

1. The most widely used estimates of variation are based on the differences, or deviations, between the estimate of location and the observed data.

$$\begin{aligned} \text{Variance} &= s^2 = \frac{\sum(x-\bar{x})^2}{n-1} \\ \text{Standard deviation} &= s = \sqrt{\text{Variance}} \end{aligned} \tag{5}$$

Contd...

1. The standard deviation is much easier to interpret than the variance since it is on the same scale as the original data.
2. Still, with its more complicated and less intuitive formula, it might seem peculiar that the standard deviation is preferred in statistics over the mean absolute deviation.
3. It owes its preeminence to statistical theory: mathematically, working with squared values is much more convenient than absolute values, especially for statistical models.

Robustness

The variance and standard deviation are especially sensitive to outliers since they are based on the squared deviations.

A robust estimate of variability is the median absolute deviation from the median or MAD:

$$Median\ absolute\ deviation = Median(|x_1 - m|, |x_2 - m|, \ldots, |x_n - m|) \tag{6}$$

where m is the median. Like the median, the MAD is not influenced by extreme values.

Estimates Based on Percentiles

1. A different approach to estimating dispersion is based on looking at the spread of the sorted data.
2. Statistics based on sorted (ranked) data are referred to as order statistics. The most basic measure is the range: the difference between the largest and smallest number.
3. The minimum and maximum values themselves are useful to know, and helpful in identifying outliers, but the range is extremely sensitive to outliers and not very useful as a general measure of dispersion in the data.

Contd...

1. To avoid the sensitivity to outliers, we can look at the range of the data after dropping values from each end.
2. Formally, these types of estimates are based on differences between percentiles. In a data set, the Pth percentile is a value such that at least P percent of the values take on this value or less and at least (100 – P) percent of the values take on this value or more. For example, to find the 80th percentile, sort the data.
3. Then, starting with the smallest value, proceed 80 percent of the way to the largest value. Note that the median is the same thing as the 50th percentile.
4. The percentile is essentially the same as a quantile, with quantiles indexed by fractions (so the .8 quantile is the same as the 80th percentile).

Contd...

1. A common measurement of variability is the difference between the 25th percentile and the 75th percentile, called the interquartile range (or IQR).
2. Here is a simple example: 3,1,5,3,6,7,2,9. We sort these to get 1,2,3,3,5,6,7,9.
3. The 25th percentile is at 2.5, and the 75th percentile is at 6.5, so the interquartile range is $6.5 - 2.5 = 4$.

Contd...

1. Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q1, Q2 and Q3, respectively.
2. Q2 is nothing but the median, since it indicates the position of the item in the list and thus, is a positional average. To find quartiles of a group of data, we have to arrange the data in ascending order.
3. In the median, we can measure the distribution with the help of lesser and higher quartile. Apart from mean and median, there are other measures in statistics, which can divide the data into specific equal parts. A median divides a series into two equal parts.

Contd...

1. Suppose, Q_3 is the upper quartile is the median of the upper half of the data set. Whereas, Q_1 is the lower quartile and median of the lower half of the data set. Q_2 is the median. Consider, we have n number of items in a data set. Then the quartiles are given by;
   Q1 = [(n+1)/4]th item
   Q2 = [(n+1)/2]th item
   Q3 = [3(n+1)/4]th item

Example

Find the quartiles of the following data: 4, 6, 7, 8, 10, 23, 34.

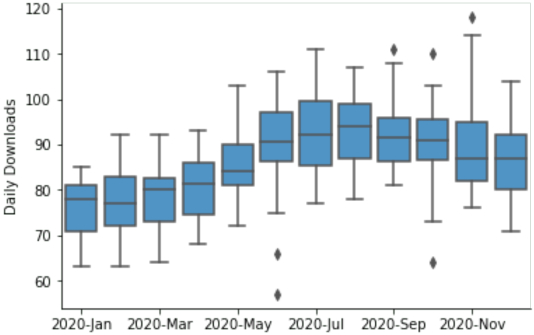
IQR

1. Interquartile Range The interquartile range (IQR) is the difference between the upper and lower quartile of a given data set and is also called a midspread.
2. It is a measure of statistical distribution, which is equal to the difference between the upper and lower quartiles. Also, it is a calculation of variation while dividing a data set into quartiles.
3. If Q_1 is the first quartile and Q_3 is the third quartile, then the IQR formula is given by

Box plot

1. A box plot (aka box and whisker plot) uses boxes and lines to depict the distributions of one or more groups of numeric data.
2. Box limits indicate the range of the central 50% of the data, with a central line marking the median value.
3. Lines extend from each box to capture the range of the remaining data, with dots placed past the line edges to indicate outliers.

Illustration



Contd...

1. The example box plot above shows daily downloads for a fictional digital app, grouped together by month.
2. From this plot, we can see that downloads increased gradually from about 75 per day in January to about 95 per day in August.
3. There also appears to be a slight decrease in median downloads in November and December.
4. Points show days with outlier download counts: there were two days in June and one day in October with low downloads compared to other days in the month.

When to use Box plots?

1. Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups.
2. They are built to provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers.
3. It is easy to see where the main bulk of the data is, and make that comparison between different groups.

Understanding Box plots

1. Construction of a box plot is based around a dataset's quartiles, or the values that divide the dataset into equal fourths.
2. The first quartile (Q1) is greater than 25% of the data and less than the other 75%. The second quartile (Q2) sits in the middle, dividing the data in half. Q2 is also known as the median.
3. The third quartile (Q3) is larger than 75% of the data, and smaller than the remaining 25%.
4. In a box and whiskers plot, the ends of the box and its center line mark the locations of these three quartiles.

Outliers from Boxplots

1. The distance between Q_3 and Q_1 is known as the interquartile range (IQR) and plays a major part in how long the whiskers extending from the box are.
2. Each whisker extends to the furthest data point in each wing that is within 1.5 times the IQR.
3. Any data point further than that distance is considered an outlier, and is marked with a dot.

Mode

1. Consider a situation in which an ice cream shop owner wants to know which flavour of ice cream is the most popular among his customers.
2. Similarly, a footwear shop owner may like to find out what design and size of shoes are in highest demand.
3. To answer this type of questions, one can use the mode which is another measure of central tendency

Contd...

1. The mode $\tilde{x}_M$ of n observations $x_1, x_2, ..., x_n$ is the value which occurs the most compared with all other values, i.e. the value which has maximum absolute frequency.
2. It may happen that two or more values occur with the same frequency in which case the mode is not uniquely defined. A formal definition of the mode is

$$\tilde{x}_N = \max(x_1, x_2, \dots, x_n) \tag{7}$$

Coefficient of Variation

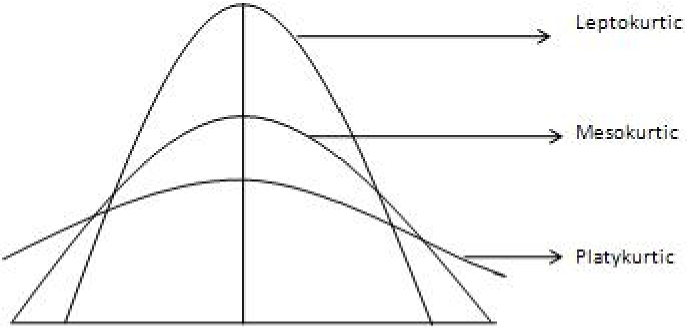
1. Consider a situation where two different variables have arithmetic means $\bar{x}_1$ and $\bar{x}_2$ with standard deviations $\tilde{SD}_1$ and $\tilde{SD}_2$, respectively.
2. Suppose we want to compare the variability of hotel prices in Munich (measured in euros) and London (measured in British pounds). How can we provide a fair comparison?
3. Coefficient of variation (CV) is given as

$$CV = \frac{\tilde{SD}}{\bar{x}} \tag{8}$$

Kurtosis

1. The degree of tailedness of a distribution is measured by kurtosis.
2. It tells us the extent to which the distribution is more or less outlier-prone (heavier or light-tailed) than the normal distribution.

Contd...



Contd...

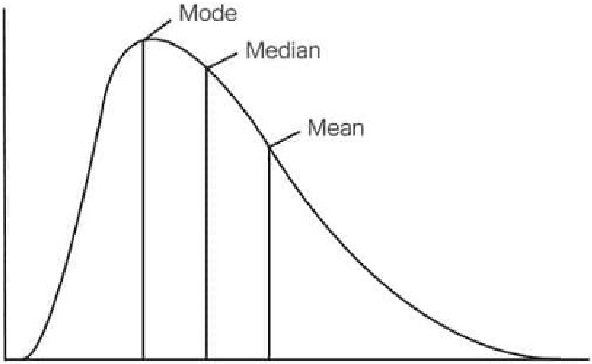
1. The data can be heavy-tailed, and the peak can be flatter, almost like punching the distribution or squishing it. This is called Negative Kurtosis (Platykurtic).
2. If the distribution is light-tailed and the top curve steeper, like pulling up the distribution, it is called Positive Kurtosis (Leptokurtic)
3. The expected value of kurtosis is 3. This is observed in a symmetric distribution. A kurtosis greater than three will indicate Positive Kurtosis.
4. In this case, the value of kurtosis will range from 1 to infinity. Further, a kurtosis less than three will mean a negative kurtosis. The range of values for a negative kurtosis is from -2 to infinity. The greater the value of kurtosis, the higher the peak.

Skewness

1. Skewness is used to measure the level of asymmetry in our graph. It is the measure of asymmetry that occurs when our data deviates from the norm.
2. Sometimes, the normal distribution tends to tilt more on one side.
3. This is because the probability of data being more or less than the mean is higher and hence makes the distribution asymmetrical. This also means that the data is not equally distributed.

Contd...

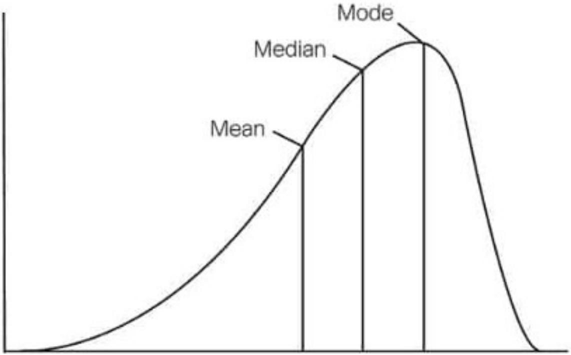
1. Positively Skewed: In a distribution that is Positively Skewed, the values are more concentrated towards the right side, and the left tail is spread out.
2. Hence, the statistical results are bent towards the left-hand side. Hence, that the mean, median, and mode are always positive. In this distribution, $Mean > Median > Mode$



Right-Skewed

Contd...

1. Negatively Skewed: In a Negatively Skewed distribution, the data points are more concentrated towards the right-hand side of the distribution. This makes the mean, median, and mode bend towards the right.
2. Hence these values are always negative. In this distribution, $Mode > Median > Mean$.



Left-Skewed (Negative Skewness)

Pearson’s First Coefficient

1. The median is always the middle value, and the mean and mode are the extremes, so you can derive a formula to capture the horizontal distance between mean and mode.

$$Pearson's\ FC = \frac{Median - Mode}{Standard\ \ Deviation} \tag{9}$$

$$Mode = 3(Median) - 2(Mean) \tag{10}$$

$$Pearson's\ SC = \frac{3(Mean - Median)}{Standard\ \ Deviation} \tag{11}$$

Chapter-II
Bivariate Analysis

Introduction

1. Correlation
2. Scatter plots
3. Coefficients for Correlation (Pearson, Karl's & Spearman)
4. Introduction to Linear Regression
5. Goodness of fit, metrics, Error analysis

Correlation

1. Exploratory data analysis in many modeling projects (whether in data science or in research) involves examining correlation among predictors, and between predictors and a target variable.
2. Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y.
3. If high values of X go with low values of Y, and vice versa, the variables are negatively correlated

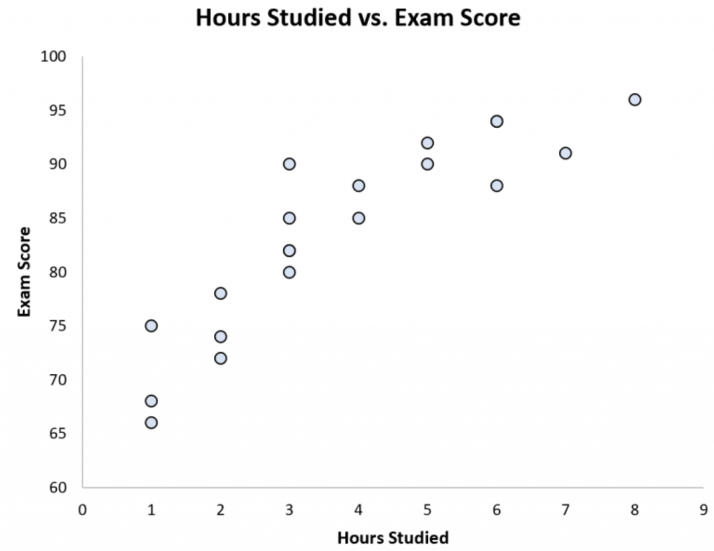
Scatter Plots

- A scatter plot is a means to represent data in a graphical format.
- A simple scatter plot makes use of the Coordinate axes to plot the points, based on their values.
- A scatter plot helps find the relationship between two variables. This relationship is referred to as a correlation. Based on the correlation, scatter plots can be classified as follows.
  a) Scatter Plot for Positive Correlation
  b) Scatter Plot for Negative Correlation
  c) Scatter Plot for Null Correlation

Why Bivariate analysis?

- Analysis of two variables simultaneously correlations, comparisons, relationships, causes, explanations tables where one variable is contingent on the values of the other variable.

Scatter Plots- Example[1]



[1] We can clearly see that there is a positive relationship between the two variables: As hours studied increases, exam score tends to increase as well.

Pearson's Correlation Coefficient

1. The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:
   - Pearson's r
   - Bivariate correlation
   - Pearson product-moment correlation coefficient (PPMCC)
   - The correlation coefficient
2. The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

Mathematical Notion

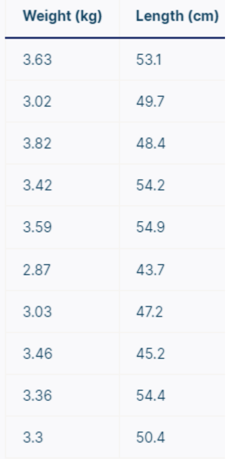
$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{12}$$

When to use Pearson's Coefficient

1. Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
2. The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
3. The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatter plot is one way to check for outliers—look for points that are far away from the others.
4. The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatter plot to check whether the relationship between two variables is linear.

Example

1. Imagine that you're studying the relationship between newborns' weight and length. You have the weights and lengths of the 10 babies born last month at your local hospital. After you convert the imperial measurements to metric, you enter the data in a table:



Weight (kg)	Length (cm)
3.63	53.1
3.02	49.7
3.82	48.4
3.42	54.2
3.59	54.9
2.87	43.7
3.03	47.2
3.46	45.2
3.36	54.4
3.3	50.4

Spearman Rank Correlation Coefficient

1. Spearman's Rank Correlation Coefficient: While calculating the correlation coefficient or product-moment correlation coefficient, it is assumed that both characteristics are measurable.
2. But, in reality, some characteristics are not measurable. For example, the qualities of individuals are not measurable. Instead, they can be ranked based on their qualities.
3. In such cases, rank correlation is used to determine the relationship between two characteristics.

Contd...

1. Sometimes there will be no clear linear relationship between two random variables, but a monotonic relationship is evident.
2. While Karl Pearson's correlation coefficient indicates the strength of a linear relationship between two variables, Spearman's rank correlation coefficient indicates the concentration of association between two qualitative characteristics.
3. In general, characteristics must be measurable to calculate the product-moment correlation coefficient. But some characteristics are not measurable in practical situations. This situation arises when dealing with qualitative studies such as honesty, beauty, and voice.

Interpretation of Spearman's Rank Correlation Coefficient[2]

1. The Spearman's rank correlation coefficient, ρ , ranges from $+1$ to 1
2. If the correlation coefficient is 1 , then all of the rankings for each variable match up for every data pair.
3. When the correlation coefficient is 1 , the rankings for one variable are the inverse of the rankings for the other variable.
4. A correlation coefficient value close to 0 indicates that the variable rankings do not have a monotonic relationship.

Summary

1. Hence, in simpler words, we can say that, ρ=1 denotes a perfect association of ranks
2. ρ=0 denotes no association between ranks
3. ρ=–1 denotes a perfect negative correlation of ranks.

[2] The closer the value of ρ is to zero, the weaker the association or correlation between the ranks.

Formula of Spearman's Rank Correlation Coefficient

1. It is given as

$$\rho = 1 - \frac{6\sum d_i^2}{n^3 - n} \tag{13}$$

2. where n is the sample size, and d_i is the difference between the ranks of the two variables $(X$ and $Y)$ under consideration.

Example: Three judges in a beauty contest assess five persons. We have to find out which pair of judges have the nearest approach to the common perception of beauty.

Manushi Chillar	1.2	2.6	6.7	4.9	5.1
Aashna Shroff	2.1	4.6	3.7	6.9	9.4
Sid Malhotra	8	2.9	6	4	7.4

Solution: We will evaluate correlation b/w Judges Manushi & Aashna

X	Rank (X)	Y	Rank (Y)	Diff	Diff^2

Table: Caption

Linear Regression in Data Science

Introduction

- Relationship between variables
- Independent and dependent variables
- One variable may or may not heavily depend on other variables

What are these variables?

- Entities such as sales, income, stock price, and weather parameters
- Difference between prediction and classification.
- Prediction involves continuous variables or range. For example: stock price
- Classification deals with discrete variables. For example, classifying cancer stages

Linear Regression: Intro

- A statistical learning technique
- Often used for predictive modeling
- Benchmark model for R&D in AI & ML

Crux

Linear regression algorithms show a linear relationship between a dependent variable, y, and one or more independent variables,x i.e., how the value of the dependent variable, y changes according to the value of the independent variable.

Contd...

- Linear regression can be of univariate or multivariate nature
- Univariate linear regression deals with a single independent variable x
- Multivariate linear regression deals with more than one independent variable $x_1, x_2, \ldots, x_n$

Univariate Linear Regression

- A firm decides to check out the relationship between advertising and sales, which can be modeled by the use of linear regression.
- The estimation of the price of a house depending on the number of rooms it has increases or decreases.
- The estimation of the price of a house depending on the number of rooms it has increases or decreases.

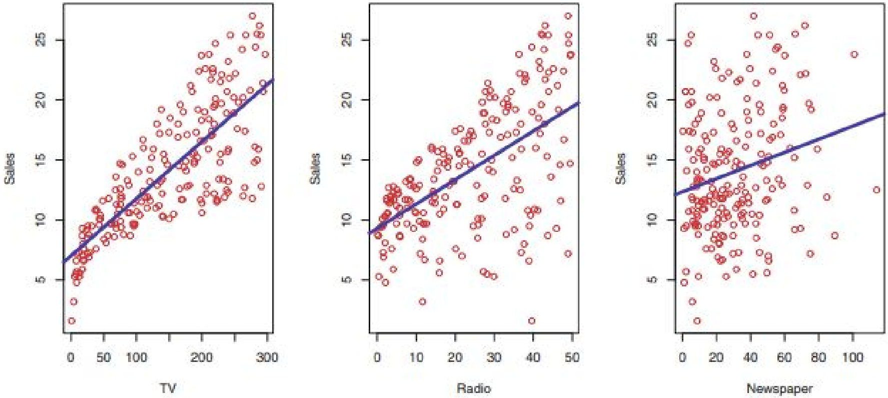
Case Study: Sales v/s Advertising Budget

- Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.

Contd...

- In this setting, the advertising budgets are input variables while sales input is an output variable.
- The input variables are typically denoted using the symbol X, with a subscript to distinguish them. So X_1 might be the TV budget, X_2 the radio budget, and X_3 the newspaper budget.
- The inputs go by different names, such as predictors, independent variables, features or predictor
- In this case, sales—is variable often called the response or dependent variable, and is typically denoted using the symbol Y

Contd...


Sales
25
20
15
10
5
0
50
100
200
300
TV
Sales
25
20
15
10
5
0
10
20
30
40
50
Radio
Sales
25
20
15
10
5
0
20
40
60
80
100
Newspaper

Important Points

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?

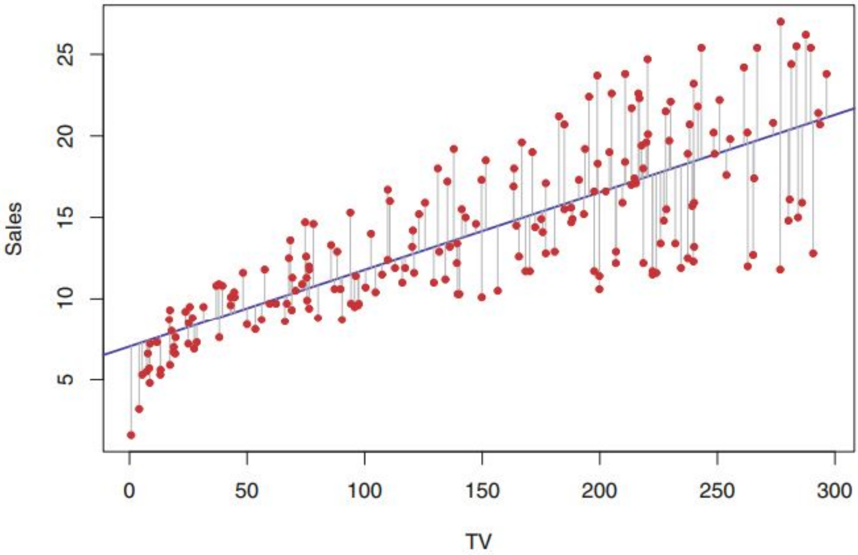
Univariate Regression for Sales v/s Advertising Budget

- Given predictor X (TV budget) and response Y (Sales), we have

$$\texttt{Sales} \approx \theta_1 \times \texttt{TV} + \theta_0 \tag{14}$$

- θ_1 and θ_0 are known as model coefficients or parameters
- Once these parameters are known from training data, we can predict future values of sales based on a given TV budget

Scatter plot for visualization



Univariate Linear Regression: Mathematical Perspective

- Consider a univariate variable x_i as input or independent variable where $i = 1, 2, \ldots, n$
- Given (x_i, y_i) as the pair of input and output variables
- A univariate linear regression model is expressed as

$$y = \theta_1 x + \theta_0 + \varepsilon \tag{15}$$

  where θ_1 is the slope parameter and θ_0 is the intercept
- θ_1 and θ_0 are known as regression parameters

Least-squares for coefficient estimation

- The linear regression model for each sample can be given as

$$y_i = \theta_1 x_i + \theta_0 + \varepsilon_i \tag{16}$$

- Consider the objective of linear regression to be $J(\theta_1, \theta_0)$ which is to be minimized. We can write

$$J(\theta_1, \theta_0) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i - \theta_0\right)^2 \tag{17}$$

Contd...

- Partial derivative of J w.r.t. θ_0 and θ_1 gives

$$\frac{\partial J}{\partial \theta_0} = -2\sum_{i=1}^{n}\left(y_i - \theta_1 x_i - \theta_0\right) \tag{18}$$

$$\frac{\partial J}{\partial \theta_1} = -2\sum_{i=1}^{n}\left(y_i - \theta_1 x_i - \theta_0\right)x_i \tag{19}$$

- Equate (18) and (19) to zero and we get

$$\theta_0 = \bar{y} - \theta_1 \bar{x} \tag{20}$$

$$\theta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{21}$$

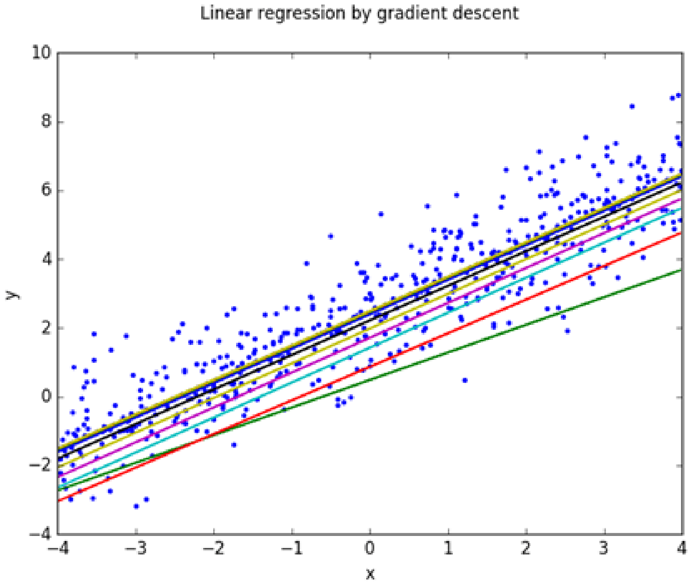
Assumptions behind Linear Regression

1. The data used in fitting the model are representative of the population.
2. The true underlying relationship between X and Y is linear.
3. The variance of the residuals is constant (homoscedastic, not heteroscedastic).
4. The residuals must be independent.
5. The residuals are normally distributed.

How to optimize parameter estimates?

- Different parameter estimates give different fit to the data
- The global aim is to minimize objective function $J(\theta_0, \theta_1)$
- We need the optimal parameters for our data
- We use gradient descent for this purpose

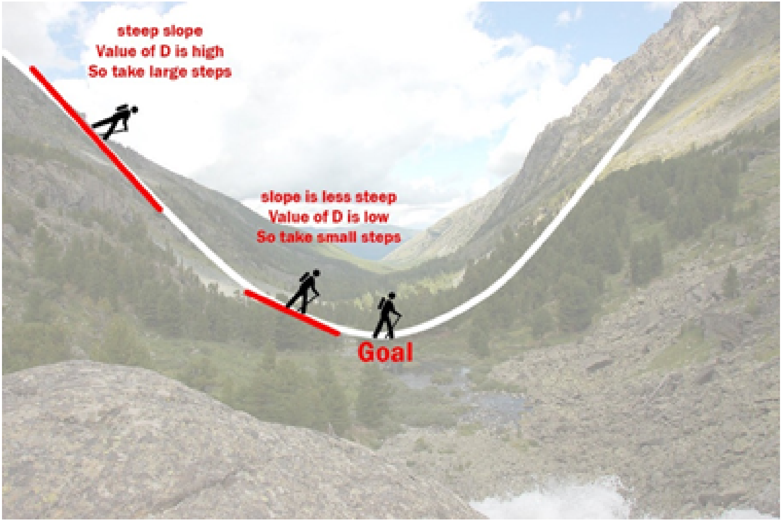
Why do we need gradient descent ?



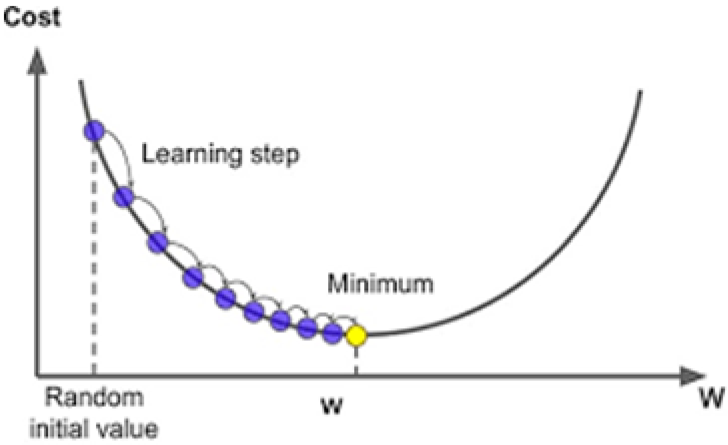
Gradient Descent for Linear Regression

- In machine learning, we use gradient descent to update the parameters of our model.
- Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient.
- Initialize your parameters with random values
-

Visual representation of gradient descent



Contd...



Gradient Descent: Mathematical Perspective

- Consider objective function J

$$\min_{\theta} \; J = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{22}$$

- In order to seek optimal parameters θ_0 and θ_1,

$$\frac{\partial J}{\partial \theta_0} = -2 \sum_{i=1}^{n} \big(y_i - \theta_1 x_i - \theta_0\big) \tag{23}$$

$$\frac{\partial J}{\partial \theta_1} = -2 \sum_{i=1}^{n} \big(y_i - \theta_1 x_i - \theta_0\big) x_i \tag{24}$$

Contd...

- Initialize $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_m]$ vector
- Update m weights (or parameters) vector using GD rule

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial J}{\partial \theta_j} \tag{25}$$

- Repeat the above step until a convergence rule is reached

$$\parallel \theta_j - \theta_{j+1} \parallel < \varepsilon \tag{26}$$

Observations

- α plays an important role in fitting
- Small value of α leads to slow convergence
- Large values of α may fail to converge or might diverge instead

Github repo

Follow this link for python implementation of gradient descent https://github.com/harshdhiman-ai/Gradient-Descent-Linear-Regression

Linear Regression in Matrix Form[3]

1. Our data consists of n paired observations of the predictor variable X and the response variable Y , i.e., $(x_1, y_1), ...(x_n, y_n)$. We wish to fit the model

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{27}$$

2. Group all of the observations of the response into a single column $(n \times 1)$ matrix y, $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

3. Similarly, we group both the coefficients into a single vector (i.e., a 2×1 matrix) $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$

[3] Please note that β_0 is equivalent to θ_0

Contd...

1. We'd also like to group the observations of the predictor variable together, but we need something which looks a little unusual at first: $\mathbf{x} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$
2. This is an $n \times 2$ matrix, where the first column is always 1, and the second column contains the actual observations of $\mathbf{X}$. We have this apparently redundant first column because of what it does for us when we multiply x by β

$$\mathbf{x}\beta = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix}$$

Minimizing Mean Squared Error

1. At each data point, using the coefficients β results in some error of prediction, so we have n prediction errors. These form a vector:

$$\epsilon(\beta) = \mathbf{y} - \mathbf{x}\beta \tag{28}$$

2. MSE is given as

$$MSE(\beta) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2(\beta)$$

3. How might we express this in terms of our matrices? I claim that the correct form is

$$MSE(\beta) = \frac{1}{n}\epsilon^T\epsilon$$

Contd...

1. To see this, look at what the matrix multiplication really involves

$$[\epsilon_1 \epsilon_2 \dots \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

2. Hence, we can rewrite the above equation as

$$\begin{aligned} MSE(\beta) &= \frac{1}{n} \epsilon^T \epsilon \\ &= \frac{1}{n} (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta) \\ &= \frac{1}{n} \left(\mathbf{y}^T - \beta^T \mathbf{x}^T\right) (\mathbf{y} - \mathbf{x}\beta) \\ &= \frac{1}{n} \left(\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{x}\beta - \beta^T \mathbf{x}^T \mathbf{y} + \beta^T \mathbf{x}^T \mathbf{x}\beta\right) \end{aligned} \tag{29}$$

Contd...

1. Consider $(y^T x\beta)^T = \beta^T x^T y$ Further notice that this is a 1×1 matrix, so Thus
2. MSE can be written as

$$MSE(\beta) = \frac{1}{n}\left(\mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{x}^T\mathbf{y} + \beta^T\mathbf{x}^T\mathbf{x}\beta\right) \tag{30}$$

3. First, we find the gradient of the MSE with respect to β:

$$\begin{aligned}\nabla MSE(\beta) &= \frac{1}{n}\left(\nabla\mathbf{y}^T\mathbf{y} - 2\nabla\beta^T\mathbf{x}^T\mathbf{y} + \nabla\beta^T\mathbf{x}^T\mathbf{x}\beta\right)\\ &= \frac{1}{n}\left(0 - 2\mathbf{x}^T\mathbf{y} + 2\mathbf{x}^T\mathbf{x}\beta\right)\\ &= \frac{2}{n}\left(\mathbf{x}^T\mathbf{x}\beta - \mathbf{x}^T\mathbf{y}\right)\end{aligned} \tag{31}$$

4. We now set this to zero at the optimum, $\hat{\beta}$:

Contd...

1. We get,

$$\mathbf{x}^T \mathbf{x} \hat{\beta} = \mathbf{x}^T \mathbf{y} \tag{32}$$

2. Hence, $\hat{\beta} = (\mathbf{x^T x})^{-\mathbf{1}} \mathbf{x^T y}$

Multivariate Linear Regression

- As the name implies, multivariate regression is a technique that estimates a single regression model with more than one outcome variable.
- When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.

Examples

- A researcher has collected data on three psychological variables, four academic variables (standardized test scores), and the type of educational program the student is in for 600 high school students.
- She is interested in how the set of psychological variables is related to the academic variables and the type of program the student is in.

Contd...

- A doctor has collected data on cholesterol, blood pressure, and weight.
- She also collected data on the eating habits of the subjects (e.g., how many ounces of red meat, fish, dairy products, and chocolate consumed per week).
- She wants to investigate the relationship between the three measures of health and eating habits.

Multivariate regression: Mathematical perspective

- Consider data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Consider m features $(x_{i1}, x_{i2}, \ldots, x_{im})$
- We estimate coefficients $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$
- The objective is to minimize residual sum of squares (RSS)

$$RSS(\theta) = \sum_{i=1}^{n} \Big(y_i - h(\theta, x_i) \Big)^2 \tag{33}$$

Contd...

- In matrix form,

$$\mathrm{RSS}(\theta) = (\boldsymbol{y} - \boldsymbol{X}\theta)^T(\boldsymbol{y} - \boldsymbol{X}\theta) \tag{34}$$

- Denote by $\boldsymbol{X}$ the $N \times (m+1)$ matrix with each row an input vector (with a 1 in the first position), and similarly let $\boldsymbol{y}$ be the N-vector of outputs in the training set.
- We have ,

$$\frac{\partial RSS}{\partial \theta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\theta) \tag{35}$$

Contd...

- Further,

$$\frac{\partial^2 RSS}{\partial\theta\partial\theta^T} = 2\boldsymbol{X}^T\boldsymbol{X} \tag{36}$$

- We set the first derivative to zero as $X^T X$ is positive definite matrix
- We get,

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\theta) = 0 \tag{37}$$

- Parameter estimates are

$$\hat{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{38}$$

Summary for Multivariate regression

- We have ,

$$Y = \theta_0 + \sum_{k=1}^{m} \theta_k X_k + \varepsilon \tag{39}$$

- ε is Gaussian random variable with zero mean and variance as σ^2
- We can express parameter estimates as

$$\hat{\theta} = \frac{\langle x, y \rangle}{\langle x, x \rangle} \tag{40}$$

Evaluating Regression Models

- The output of a regression ML model is a numeric value for the model's prediction of the target.
- For example, if you are predicting housing prices, the prediction of the model could be a value such as 254,013.
- Consider actual response to be y_i and predicted response as $\hat{y}_i$, the mean absolute error is given as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{41}$$

Contd...

- The mean squared error (MSE) is given as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{42}$$

- A popular metric root mean squared error (RMSE) is given as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{43}$$

Contd...

- Mean absolute percentage error (MAPE) is given as

$$\mathrm{MAPE} = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{44}$$

- Coefficient of regression (R^2) is given as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{45}$$

Importance of R^2

- R^2 is a statistic that will give some information about the goodness of fit of a model.
- In regression, the R^2 coefficient of determination is a statistical measure of how well the regression predictions approximate the real data points.
- An R^2 of 1 indicates that the regression predictions perfectly fit the data.

Module-III
Sampling Techniques

Introduction

The Idea of Sampling

1. Idea 1: Examine a part of a whole
2. Idea 2: Randomize
3. Idea 3: It's the sample size

Example

1. In 1936, a young pollster named George Gallup used a subsample of only 3000 of the 2.4 million responses that the Literary Digest received to reproduce the wrong prediction of Landon's victory over Roosevelt.
2. He then used an entirely different sample of 50,000 and predicted that Roosevelt would get 56% of the vote to Landon's 44%.
3. His sample was apparently much more representative of the actual voting populace. The Gallup Organization went on to become one of the leading polling companies.

Idea I

1. The first idea is to draw a sample. We'd like to know about an entire group of individuals—a population—but examining all of them is usually impractical, if not impossible.
2. So we settle for examining a smaller group of individuals—a sample—selected from the population.

Contd...

1. You do this every day. For example, suppose you wonder how the vegetable soup you're cooking for dinner tonight is going to go over with your friends.
2. To decide whether it meets your standards, you only need to try a small amount.
3. You might taste just a spoonful or two. You certainly don't have to consume the whole pot.
4. You trust that the taste will represent the flavor of the entire pot. The idea behind your tasting is that a small sample, if selected properly, can represent the entire population.

Bias in Sampling

1. Selecting a sample to represent the population fairly is more difficult than it sounds.
2. Polls or surveys most often fail because they use a sampling method that tends to over- or under-represent parts of the population
3. Sampling methods that, by their nature, tend to over- or underemphasize some characteristics of the population are said to be biased.
4. Bias is the bane of sampling—the one thing above all to avoid. Conclusions based on samples drawn with biased methods are inherently flawed. There is usually no way to fix bias after the sample is drawn and no way to salvage useful information from it.

Idea 2: Randomize

1. Think back to the soup sample. Suppose you add some salt to the pot. If you sample it from the top before stirring, you'll get the misleading idea that the whole pot is salty.
2. If you sample from the bottom, you'll get an equally misleading idea that the whole pot is bland.
3. By stirring, you randomize the amount of salt throughout the pot, making each taste more typical of the whole pot.

Idea 3: It's the sample size

1. How large a random sample do we need for the sample to be reasonably representative of the population?
2. Most people think that we need a large percentage, or fraction, of the population, but it turns out that all that matters is the number of individuals in the sample.

Contd...

1. How big a sample do you need? That depends on what you're estimating. To get an idea of what's really in the soup, you'll need a large enough taste to get a representative sample from the pot.
2. For a survey that tries to find the proportion of the population falling into a category, you'll usually need several hundred respondents to say anything precise enough to be useful

Populations & parameters

- A study found that teens were less likely to “buckle up.” The National Center for Chronic Disease Prevention and Health Promotion reports that 21.7% of U.S. teens never or rarely wear seat belts.
- We’re sure they didn’t take a census, so what does the 21.7% mean? We can’t know what percentage of teenagers wear seat belts. Reality is just too complex. But we can simplify the question by building a model.

Contd...

- Models use mathematics to represent reality.
- Parameters are the key numbers in those models. A parameter used in a model for a population is sometimes called (redundantly) a population parameter.
- But let's not forget about the data. We use summaries of the data to estimate the population parameters.
- As we know, any summary found from the data is a statistic. Sometimes you'll see the (also redundant) term sample statistic

Contd...

1. We draw samples because we can't work with the entire population, but we want the statistics we compute from a sample to reflect the corresponding parameters accurately.
2. A sample that does this is said to be representative.
3. A biased sampling methodology tends to over- or underestimate the parameter of interest

Exercise: Various claims are often made for surveys. Why is each of the following claims not correct?

1. Stopping students on their way out of the cafeteria is a good way to sample if we want to know about the quality of the food there.
2. We drew a sample of 100 from the 3000 students in a school. To get the same level of precision for a town of 30,000 residents, we'll need a sample of 1000.

Sampling Methods

1. How would you select a representative sample? Most people would say that every individual in the population should have an equal chance to be selected, and certainly that seems fair.
2. But it's not sufficient. There are many ways to give everyone an equal chance that still wouldn't give a representative sample.

Example

1. Consider, for example, a school that has equal numbers of males and females. We could sample like this: Flip a coin.
2. If it comes up heads, select 100 female students at random. If it comes up tails, select 100 males at random.
3. Everyone has an equal chance of selection, but every sample is of only a single gender—hardly representative

Simple Random Sampling

1. We need to do better. Suppose we insist that every possible sample of the size we plan to draw has an equal chance to be selected. This ensures that situations like the one just described are not likely to occur and still guarantees that each person has an equal chance of being selected.
2. What's different is that with this method, each combination of people has an equal chance of being selected as well.
3. A sample drawn in this way is called a Simple Random Sample, usually abbreviated SRS.

Contd...

1. To select a sample at random, we first need to define where the sample will come from. The sampling frame is a list of individuals from which the sample is drawn.
2. For example, to draw a random sample of students at a college, we might obtain a list of all registered full-time students and sample from that list

Contd...

1. Samples drawn at random generally differ one from another.
2. Each draw of random numbers selects different people for our sample. These differences lead to different values for the variables we measure.
3. We call these sample-to-sample differences sampling variability.

Method 2: Stratified Sampling

1. Designs that are used to sample from large populations—especially populations residing across large areas—are often more complicated than simple random samples.
2. Sometimes the population is first sliced into homogeneous groups, called strata, before the sample is selected. Then simple random sampling is used within each stratum before the results are combined.
3. This common sampling design is called stratified random sampling.

Example

Situation

You’re trying to find out what freshmen think of the food served on campus. Food Services believes that men and women typically have different opinions about the importance of the salad bar.

Question

How should you adjust your sampling strategy to allow for this difference?

Answer

I will stratify my sample by drawing an SRS of men and a separate SRS of women—assuming that the data from the registrar include information about each person’s gender.

Method 3: Cluster and Multistage Sampling

1. Suppose we wanted to assess the reading level of a textbook based on the length of the sentences.
2. Simple random sampling could be awkward; we'd have to number each sentence, then find, for example, the 576th sentence or the 2482nd sentence, and so on.
3. It would be much easier to pick a few pages at random and count the lengths of the sentences on those pages.
4. That works if we believe that each page is representative of the entire book in terms of reading level.

Contd...

1. Splitting the population into representative clusters can make sampling more practical.
2. Then we could simply select one or a few clusters at random and perform a census within each of them. This sampling design is called cluster sampling.
3. Clusters are generally selected for reasons of efficiency, practicality, or cost. Ideally, if each cluster represents the full population fairly, cluster sampling will be unbiased.

Contd...

Situation

In trying to find out what freshmen think about the food served on campus, you've considered both an SRS and a stratified sample. Now you have run into a problem: It's simply too difficult and time consuming to track down the individuals whose names were chosen for your sample. Fortunately, freshmen at your school are all housed in 10 freshman dorms

Question

How could you use this fact to draw a cluster sample? How might that alleviate the problem? What concerns do you have?

Answer

To draw a cluster sample, I would select a few freshman dorms at random and then try to contact everyone in each selected dorm. I could save time by simply knocking on doors on a given evening and interviewing people. I'd have to assume that freshmen were assigned to dorms pretty much at random and that the people I'm able to contact are representative of everyone in the dorm.

Stratified vs Cluster Sampling

1. Boston cream pie consists of a layer of yellow cake, a layer of custard, another cake layer, and then a chocolate frosting. Suppose you are a professional taster (yes, there really are such people) whose job is to check your company's pies for quality. You'd need to eat small samples of randomly selected pies, tasting all three components: the cake, the custard, and the frosting.

Contd...

1. One approach is to cut a thin vertical slice out of the pie. Such a slice will be a lot like the entire pie, so by eating that slice, you'll learn about the whole pie. This vertical slice containing all the different ingredients in the pie would be a cluster sample.
2. Another approach is to sample in strata: Select some tastes of the cake at random, some tastes of custard at random, and some bits of frosting at random. You'll end up with a reliable judgment of the pie's quality. Many populations you might want to learn about are like this Boston cream pie. You can think of the subpopulations of interest as horizontal strata, like the layers of pie.
3. Cluster samples slice vertically across the layers to obtain clusters, each of which may have parts of the entire population. Stratified samples represent the population by drawing some from each layer, reducing variability in the results that could arise because of the differences among the layers.

Exercise

1. We need to survey a random sample of the 300 passengers on a flight from San Francisco to Tokyo. Name each sampling method described below.

a) Pick every 10th passenger as people board the plane.

b) From the boarding list, randomly choose 5 people flying first class and 25 of the other passengers.

c) Randomly generate 30 seat numbers and survey the passengers who sit there.

d) Randomly select a seat position (right window, right center, right aisle, etc.) and survey all the passengers sitting in those seats

Sampling Mean & Sampling Distribution

Sampling Mean

1. . Given a random sample $Y_1, Y_2, \ldots, Y_n$, the sample mean is defined to be

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{46}$$

2. If n is large than there are theorems that say that However, $\bar{Y}$ is usually close to μ[4].
3. Y is a random variable.
4. Generate another random sample and we will get a different value for $\bar{Y}$.

[4] μ is known as population mean

Contd...

1. Similarly, sampling variance can be defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{47}$$

Sampling Distribution

1. A sampling distribution of a statistic is a type of probability distribution created by drawing many random samples of a given size from the same population.
2. These distributions help you understand how a sample statistic varies from sample to sample.

Contd...

1. Sampling distributions are essential for inferential statistics because they allow you to understand a specific sample statistic in the broader context of other possible values.
2. Crucially, they let you calculate probabilities associated with your sample.

Contd...

1. Theorem: Let $\bar{X}$ be the sample mean of a random sample of size n drawn from a population having mean μ and standard deviation σ, then the mean of $\bar{X}$ is

$$\mu_{\bar{X}} = \mu \tag{48}$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{49}$$

Example

The mean and standard deviation of the strength of a packaging material are 55 kg and 6 kg, respectively. A quality manager takes a random sample of specimens of this material and tests their strength. If the manager wants to reduce the standard deviation of X to 1.5 kg, how many specimens should be tested?

Central Limit Theorem

1. In probability theory, the central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger, assuming that all samples are identical in size, and regardless of the population’s actual distribution shape.
2. Put another way, CLT is a statistical premise that, given a sufficiently large sample size from a population with a finite level of variance, the mean of all sampled variables from the same population will be approximately equal to the mean of the whole population.

Key Components of CLT

1. Sampling is successive. This means some sample units are common with sample units selected on previous occasions.
2. Sampling is random. All samples must be selected at random so that they have the same statistical possibility of being selected.
3. Samples should be independent. The selections or results from one sample should have no bearing on future samples or other sample results.
4. Samples should be limited. It's often cited that a sample should be no more than 10% of a population if sampling is done without replacement. In general, larger population sizes warrant the use of larger sample sizes.
5. Sample size is increasing. The central limit theorem is relevant as more samples are selected.

Z-Score

1. Z-score is a statistical measurement that describes a value's relationship to the mean of a group of values.
2. Z-score is measured in terms of standard deviations from the mean.
3. If a Z-score is 0, it indicates that the data point's score is identical to the mean score.

Contd...

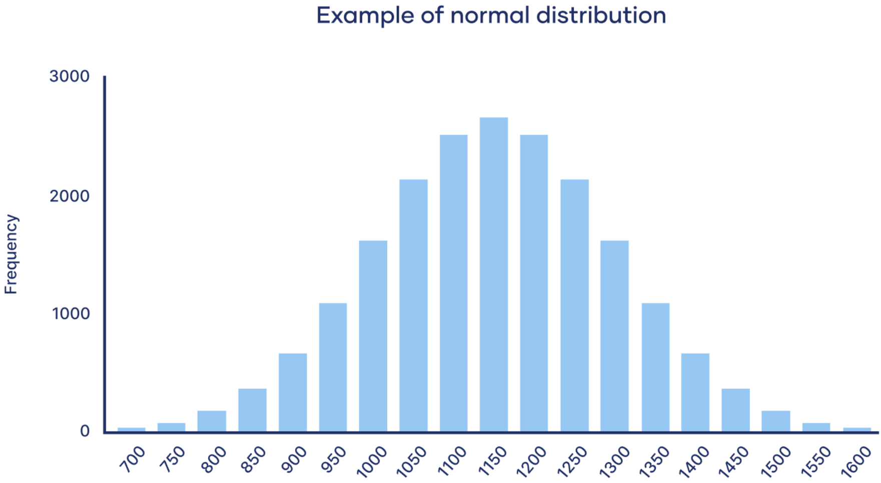
1. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean.
2. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.
3. Z-Score is computed as

$$z = \frac{x - \mu}{\sigma}, \tag{50}$$

where x is individual present in the given sample, μ is mean, and σ is standard deviation

Normal Distribution

1. In a normal distribution, data is symmetrically distributed with no skew.
2. When plotted on a graph, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.
3. Normal distributions are also called Gaussian distributions or bell curves because of their shape.

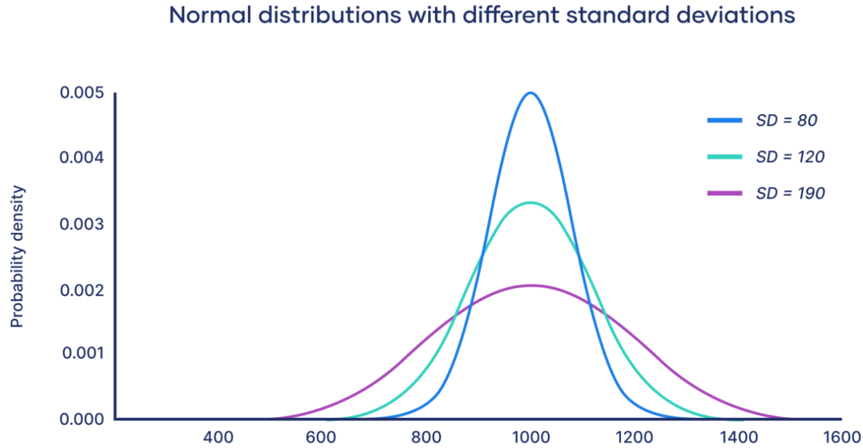

Example of normal distribution
Frequency
3000
2000
1000
0
700
750
800
850
900
950
1000
1050
1100
1150
1200
1250
1300
1350
1400
1450
1500
1550
1600

Properties of Normal Distribution

1. The mean, median and mode are exactly the same.
2. The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
3. The distribution can be described by two values: the mean and the standard deviation.

Contd...

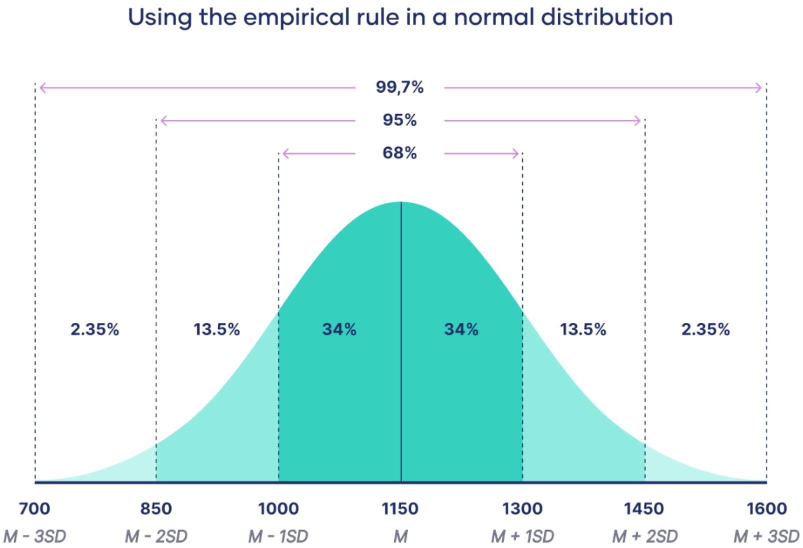
1. The mean is the location parameter while the standard deviation is the scale parameter.
2. The mean determines where the peak of the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.


Normal distributions with different standard deviations
Probability density
0.005
0.004
0.003
0.002
0.001
0.000
400
600
800
1000
1200
1400
1600
SD = 80
SD = 120
SD = 190

Empirical Rule

1. The empirical rule, or the 68-95-99.7 rule, tells you where most of your values lie in a normal distribution:
2. Around 68% of values are within 1 standard deviation from the mean.
3. Around 95% of values are within 2 standard deviations from the mean.
4. Around 99.7% of values are within 3 standard deviations from the mean.

Contd...


Using the empirical rule in a normal distribution
99,7%
95%
68%
2.35%
13.5%
34%
34%
13.5%
2.35%
700
M - 3SD
850
M - 2SD
1000
M - 1SD
1150
M
1300
M + 1SD
1450
M + 2SD
1600
M + 3SD

Probability Density Function

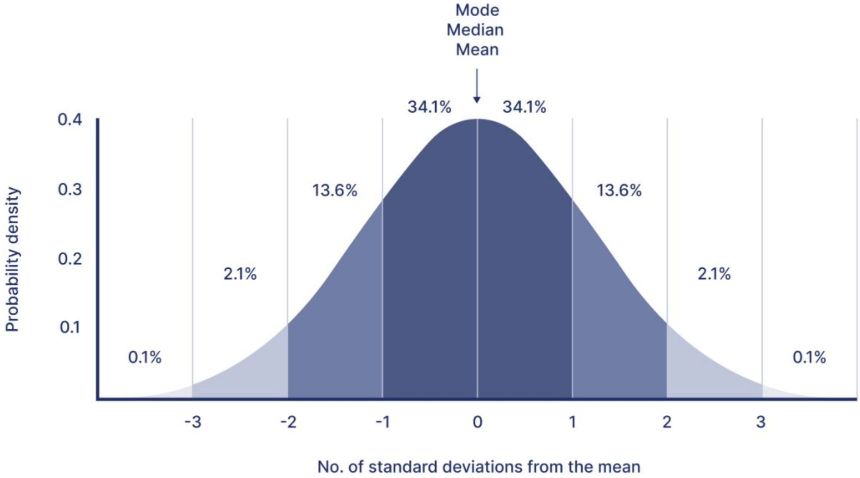
1. Given a datapoint x, mean μ, and standard deviation σ, the PDF for a normal distribution can be given as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{51}$$

Standard Normal Distribution

1. The standard normal distribution, also called the z-distribution, is a special normal distribution where the mean is 0 and the standard deviation is 1.
2. Every normal distribution is a version of the standard normal distribution that's been stretched or squeezed and moved horizontally right or left.

Z-Distribution



Finding probability using the z-distribution

1. Each z-score is associated with a probability, or p-value, that tells you the likelihood of values below that z-score occurring.
2. If you convert an individual value into a z-score, you can then find the probability of all values up to that value occurring in a normal distribution.

Z-tables

1. URL for Positive Z-table[5]
2. URL for Negative Z-table[6]

[5] https://www.ztable.net/wp-content/uploads/2018/11/positiveztable.png

[6] https://www.ztable.net/wp-content/uploads/2018/11/negativeztable.png

Example: Consider a population of GMAT scores with mean marks of 1150 and a standard deviation of 200. Determine % of students scoring more than half of the mean marks.

1. Given $\mu = 1150$ and $\sigma = 200$ and $x = 575$
2. Compute z-score for $x = 575$ as

$$
\begin{aligned}
Z &= \frac{x-\mu}{\sigma} \quad (52)\\
&= \frac{575-1150}{200} \quad (53)\\
&= -2.875 \quad (54)
\end{aligned}
$$

3. Look for p-value in Z-table for a Z-score of -2.875
4. We get, p =0.00205, i.e. the probability of SAT score less than 575 is 0.21%
5. Thus, probability of score more than 575 is 99.79%

Module-IV
Inferential Statistics: Hypothesis Testing

Introduction

1. Statistical hypothesis: a claim about the value of a parameter or population characteristic.
2. To prove that a hypothesis is true, or false, with absolute certainty, we would need absolute knowledge. That is, we would have to examine the entire population.
3. Instead, hypothesis testing concerns on how to use a random sample to judge if it is evidence that supports or not the hypothesis.

Contd...

1. Hypothesis testing is formulated in terms of two hypotheses:
   (1) H_0: the null hypothesis;
   (2) H_1: the alternate hypothesis.
2. In hypothesis testing, two mutually exclusive statements about a parameter or population (hypotheses) are evaluated to decide which statement is best supported by sample data.

Contd...

1. Hypothesis tests including a specific parameter are called parametric tests.
2. In parametric tests, the population is assumed to have a normal distribution (e.g., the height of people in a class).
3. In contrast, non-parametric tests (also distribution-free tests) are used when parameters of a population cannot be assumed to be normally distributed.
4. For example, the price of diamonds may be exponentially distributed. Non-parametric doesn't mean that you do not know anything about a population but rather that it is not normally distributed.

Summary

1. In reviewing hypothesis tests, we start first with the general idea.
2. Then, we keep returning to the basic procedures of hypothesis testing, each time adding a little more detail. The general idea of hypothesis testing involves:
    - Making an initial assumption.
    - Collecting evidence (data).
    - Based on the available evidence (data), deciding whether to reject or not reject the initial assumption.
3. Every hypothesis test — regardless of the population parameter involved — requires the above three steps.

Real-World Example

1. Hypothesis: Average order value has increased since last financial year
2. Parameter: Mean order value
3. Test type: one-sample, parametric test (assuming the order value follows a normal distribution)

Contd...

1. Hypothesis: Investing in stock A brings a higher return than investing in stock B
2. Parameter: Difference in mean return
3. Test type: two-sample, parametric test, also AB test (assuming the return follows a normal distribution)

Sample Tests

1. One-sample, two-sample, or more-sample test When testing hypotheses, it is distinguished between one-sample, two-sample or more-sample tests.
2. In a one-sample test, a sample (average order value this year) is compared to a known value (average order value of last year).
3. In a two-sample test, two samples (investment A and B) are compared to each other.

Null & Alternative Hypothesis

1. The null and alternative hypotheses are the two mutually exclusive statements about a parameter or population
2. The null hypothesis (often abbreviated as H0) claims that there is no effect or no difference.
3. The alternative hypothesis (often abbreviated as H1 or HA) is what you want to prove. Using one of the examples from above:
   - H0: There is no difference in the mean return from A and B, or the difference between A and B is zero.
   - H1: There is a difference in the mean return from A and B or the difference between A and B $>$ zero.

One-sided and two-sided (one-tailed and two-tailed) tests

1. In a two-tailed test, you are testing in both directions, meaning it is tested whether the mean return from A is significantly greater and significantly less than the mean return from B.
2. In a one-tailed test, you are testing in one direction, meaning it is tested either if the mean return from A is significantly greater or significantly less than the mean return from B.
3. In this case, the alternative hypothesis would change to either of:
   - H1: The mean return of A is greater than the mean return of B.
   - H1: The mean return of A is lower than the mean return of B.

Selection of Test Statistic

1. To test your claims, you need to decide on the right test or test statistic.
2. Often discussed tests are the t-test, z-test, or F-test, which all assume a normal distribution.
3. However, in business, a normal distribution often cannot be assumed.

Contd...

1. Parametric or non-parametric test, each test has a test statistic. A test statistic is a numerical summary of a sample.
2. It is a random variable as it is derived from a random sample.
3. In hypothesis tests, it compares the sample statistic to the expected result of the null hypothesis. The selection of the test statistic is dependent on:
    - Parametric vs. non-parametric
    - Number of samples (one, two, multiple)
    - Discrete (e.g. number of customers) or continuous variable (e.g. order value)

One-Sample Z-test

1. A one sample mean test is used when the population is known to be normally distributed and when the population standard deviation (σ) is known.
2. The formula for computing a test statistic for one sample mean is identical to that of computing a test statistic for one sample mean, except now the population standard deviation is known and can be used in computing the standard error.

Z-statistic

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{55}$$

- $\bar{x}$ is known as sample mean
- μ_0 is known as hypothesized population mean
- n is the sample size under consideration
- σ is population standard deviation

Understanding Level of Significance

1. The level of significance is the measurement of the statistical significance.
2. It defines whether the null hypothesis is assumed to be accepted or rejected.
3. It is expected to identify if the result is statistically significant for the null hypothesis to be false or rejected.

Contd...

1. The level of significance is denoted by the Greek symbol α (alpha). Therefore, the level of significance is defined as follows:
2. Significance Level = p (type I error) = α

Contd...

1. Example: The value significant at 5% refers to p-value is less than 0.05 or $p < 0.05$.
2. Similarly, significant at the 1% means that the p-value is less than 0.01.

Summary

1. The level of significance is taken at 0.05 or 5%.
2. When the p-value is low, it means that the recognised values are significantly different from the population value that was hypothesised in the beginning.
3. The p-value is said to be more significant if it is as low as possible. Also, the result would be highly significant if the p-value is very less.
4. But, most generally, p-values smaller than 0.05 are known as significant, since getting a p-value less than 0.05 is quite a less practice.

How to Find the Level of Significance?

1. To measure the level of statistical significance of the result, the investigator first needs to calculate the p-value.
2. It defines the probability of identifying an effect which provides that the null hypothesis is true.
3. When the p-value is less than the level of significance (α), the null hypothesis is rejected. If the p-value so observed is not less than the significance level α, then theoretically null hypothesis is accepted.
4. But practically, we often increase the size of the sample size and check if we reach the significance level.

Generic Interpretation

1. If $p > 0.1$, then there will be no assumption for the null hypothesis
2. If $p > 0.05$ and $p \leq 0.1$, it means that there will be a low assumption for the null hypothesis.
3. If $p > 0.01$ and $p \leq 0.05$, then there must be a strong assumption about the null hypothesis.
4. If $p < 0.01$, then a very strong assumption about the null hypothesis is indicated.

Example for Level of Significance

1. If we obtain a p-value equal to 0.03, then it indicates that there are just 3% chances of getting a difference larger than that in our research, given that the null hypothesis exists.
2. Now, we need to determine if this result is statistically significant enough.
3. We know that if the chances are 5% or less than that, then the null hypothesis is true, and we will tend to reject the null hypothesis and accept the alternative hypothesis.
4. Here, in this case, the chances are 0.03, i.e. 3% (less than 5%), which eventually means that we will eliminate our null hypothesis and will accept an alternative hypothesis.

Example

1. The population of all verbal GRE scores are known to have a standard deviation of 8.5. The UWPsychology department hopes to receive applicants with a verbal GRE scores over 210. This year, the mean verbal GRE scores for the 42 applicants was 212.79. Using a value of $\alpha = 0.05$ is this new mean significantly greater than the desired mean of 210?
2. Formulate Null Hypothesis as $H_0 : \mu = \mu_0 = 210$
3. Alternate Hypothesis $H_1 : \mu > \mu_0$

Contd...

1. Given the sample mean $\bar{x} = 212.79$, with $\mu = 8.5$
2. We have a sample size of $n = 42$
3. Compute z-statistic as
$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{56}$$
4. Hence $z = 2.13$
5. This will be a one tailed test because we're only rejecting H_0 if our observed mean is significantly larger than 210.
6. To make our decision we need to find the critical value of z, which is the z for which the area above is 0.05. Looking at our z-table for $\alpha = 0.05$

Contd...

1. Our observed value of z is 2.13 which is greater than the critical value of 1.64.
2. We therefore reject H_0.

Example: Two-tailed test

1. Suppose you start up a company that has developed a drug that is supposed to increase IQ.
2. You know that the standard deviation of IQ in the general population is 15. You test your drug on 36 patients and obtain a mean IQ of 97.65.
3. Using an alpha value of 0.05, is this IQ significantly different than the population mean of 100?

Solution

1. First, we calculate the standard error as

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{57}$$

2. We get $\sigma_{\bar{x}} = 2.5$
3. Next, calculate the z-statistic as

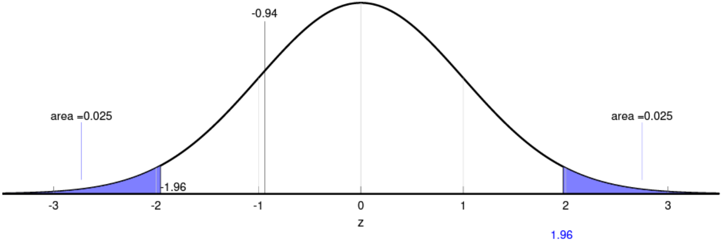
$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{58}$$

4. We get, $z = -0.94$

Contd...

1. We then compare our observed value of z to the critical values of z for alpha = 0.05. We are looking for a significant difference, so this will be a two-tailed test.
2. We reject the null hypothesis if our observed mean is either significantly larger or smaller than 100.
3. Our critical values of z are therefore the two values that span the middle 95% of the area under the standard normal distribution. This means that the areas in each of the two tails is 0.05 = 0.025

Visualizing Rejection Region



Contd...

1. The rejection region contains values of z less than-1.96 and greater than 1.96.
2. Our observed value of z falls outside the rejection region, so we fail to reject H0 and conclude that our drug did not have a significant effect on IQ.

Homework Problem 1

1. Suppose the jewelry of exams has a population that is normally distributed with a standard deviation of 5. You are walking down the street and sample 9 exams from this population and obtain a mean jewelry of 28.95 and a standard deviation of 6.3802. Using an alpha value of alpha = 0.01, is this observed mean significantly different than an expected jewelry of 27?

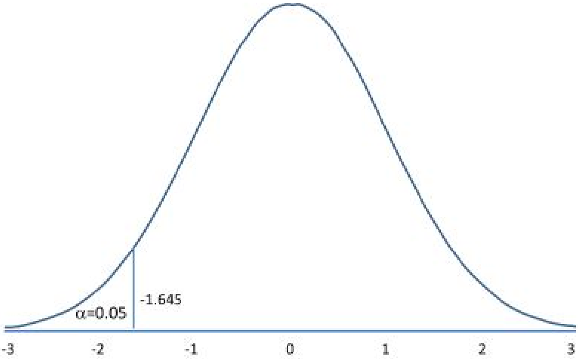
Homework Problem 2

1. Suppose the courage of psychologists has a population that is normally distributed with a standard deviation of 10. You decide to sample 57 psychologists from this population and obtain a mean courage of 34.81 and a standard deviation of 9.0579. Using an alpha value of 0.05, is this observed mean significantly greater than an expected courage of 34?

Homework Problem 3

1. Suppose the life expectancy of Seattleites has a population that is normally distributed with a standard deviation of 1. You go out and sample 45 Seattleites from this population and obtain a mean life expectancy of 88.51 and a standard deviation of 1.0815. Using an alpha value of 0.05, is this observed mean significantly different than an expected life expectancy of 89?

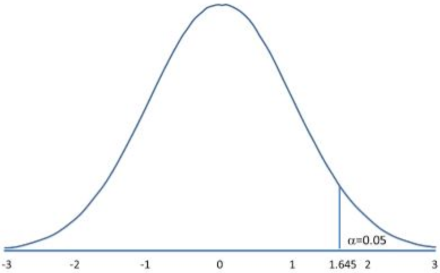
Summary: Left Tailed



Rejection Region for Lower-Tailed Z Test (H_1: $\mu < \mu_0$) with α =0.05

The decision rule is: Reject H_0 if $Z \leq 1.645$.

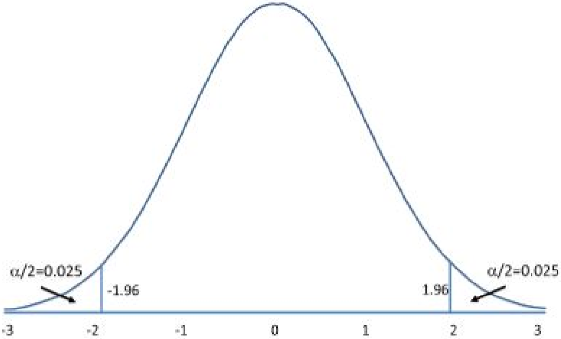
Summary: Right-tailed



Rejection Region for Upper-Tailed Z Test (H_1: $\mu > \mu_0$) with α=0.05

The decision rule is: Reject H_0 if $Z \geq 1.645$.

Summary: Two-tailed



Rejection Region for Two-Tailed Z Test (H_1: $\mu \neq \mu_0$) with $\alpha = 0.05$

The decision rule is: Reject H_0 if $Z \leq -1.960$ or if $Z \geq 1.960$.

Example: Test the research hypothesis that the mean weight in men in 2006 is more than 191 pounds. We will assume the sample data are as follows: $n = 100$, $\bar{x}$=197.1 and σ=25.6.

1. Step 1. Set up hypotheses and determine level of significance H0: $\mu = 191$ and H1: $\mu >$191 at $\alpha = 0.05$
2. Step 2. Select the appropriate test statistic. Because the sample size is large ($n > 30$) the appropriate test statistic is

$$z_{obs} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{59}$$

3. Set up decision rule. In this example, we are performing an upper tailed test (H1: $\mu > 191$), with a Z test statistic and selected $\alpha = 0.05$. Reject H0 if $Z >$1.645.
4. Calculate z-statistic as

Example: Suppose the amount of beer has a population that is normally distributed with a standard deviation of 9. You are walking down the street and sample 67 beer from this population and obtain a mean amount of -0.54 and a standard deviation of 9.9197. Using an alpha value of alpha= 0.05, is this observed mean significantly less than an expected amount of 0?

Solution

1. Step 1: Frame NULL and Alternate Hypothesis as H0: $\mu = 0$ and H1: $\mu < 0$
2. Given the data, select appropriate test statistic
3. Given $\bar{x} = -0.54$ and $\sigma = 9.9197$, and $n = 67$
4. Calculate Z-statistic as

$$z_{obs} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{60}$$

5. We get,

$$z_{obs} = \frac{-0.54 - 0}{\frac{9}{\sqrt{67}}} \tag{61}$$

6. z_{crit}=-1.64 (lef-tailed test) for alpha=0.05

t-Test in Statistics

1. The t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis.
2. It can be used to determine if two sets of data are significantly different from each other.
3. It 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.

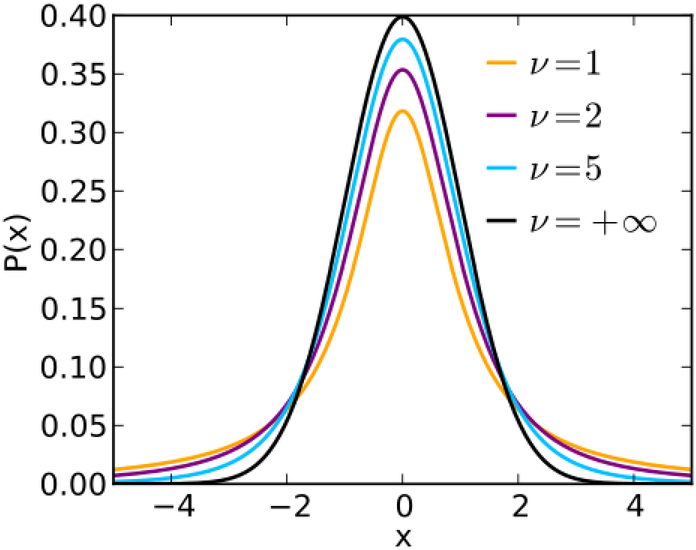
t-distribution

1. In probability and statistics, Student's t-distribution (or simply the t-distribution) is any member of a family of continuous probability distributions that arise when estimating the mean of a normally distributed population in situations where the sample size is small and the population's standard deviation is unknown.
2. It was developed by English statistician William Sealy Gosset under the pseudonym "Student".

Contd...

1. The t-distribution is symmetric and bell-shaped, like the normal distribution.
2. However, the t-distribution has heavier tails, meaning that it is more prone to producing values that fall far from its mean.

t-distribution



Contd...

1. Let $X_1, X_2, \ldots, X_n$ denote the independently and identically drawn from a population $N(\mu, \sigma^2)$ is a sample of size n from a normally distributed population with expected mean of μ and variance σ^2
2. t-test is used for two cases (a) Test for single mean (b) Test for paired mean

t-Test for single mean

1. Form null hypothesis H_0, there is no significance difference between the sample mean and the population mean ie., $\mu = \mu_0$
2. Form alternate hypothesis $H_1 : \mu \leq \mu_0, \mu > \mu_0$. There is significance difference between the sample mean and the population mean
3. Level of significance: Set to 5% or 1%
4. Test-statistic

$$t_{cal} = \left| \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \right| \tag{62}$$

5. Where $\bar{x}$ is sample mean and s is standard deviation of the sample

Contd...

1. Find out t-value for the given level of significance and at $n - 1$ degrees of freedom
2. Compare the calculated t-value with the t-value for the given level of significance
3. Now, follow the rule
    - If $t_{cal} < t_{tab}$, Accept H_0
    - If $t_{cal} > t_{tab}$, Accept H_1

Example

Suppose a sample of 16 light trucks is randomly selected off the assembly line. The trucks are driven 1000 miles and the fuel mileage (MPG) of each truck is recorded. It is found that the mean MPG is 22 with a SD equal to 3. The previous model of the light truck got 20 MPG.

- State the null hypothesis for the problem above
- Conduct a test of the null hypothesis at $p = .05$. BE SURE TO PROPERLY STATE YOUR STATISTICAL CONCLUSION.
- Provide an interpretation of your statistical conclusion using the variables from the description given

Solution

1. Null hypothesis can be stated as "We expect the sample of 16 light trucks to get the same average MPG as the previous model. Any observed difference in MPG between the new light trucks and the previous model is assumed to be solely due to random error."
2. Compute t-statistic, Sample Mean = 22 MPG, SD = 3 MPG, n=16 Population Mean = 20 MPG

$$t_{cal} = \left| \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \right| \tag{63}$$

$$= \frac{22 - 20}{\frac{3}{\sqrt{16}}} \tag{64}$$

$$= 2.67 \tag{65}$$

3. Compare critical t-value with calculated one. t_{crit} at 2.5% significance level and at 15 DOF is 2.131

Contd...

1. Since $t_{cal} > t_{crit}$, we reject H_0
2. The observed difference between average MPG for the new model truck and average MPG for the previous model truck is not solely due to random error ($p = .05$).
3. It appears that, on the average, the new model light truck gets slightly better gas mileage compared to the previous model ($p = .05$).

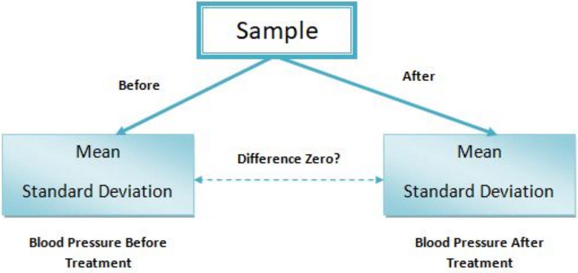
Example 2

1. We have the potato yield from 12 different farms. We know that the standard potato yield for the given variety is μ=20.
2. x = [21.5, 24.5, 18.5, 17.2, 14.5, 23.2, 22.1, 20.5, 19.4, 18.1, 24.1, 18.5]
3. Test if the potato yield from these farms is significantly better than the standard yield. (Take level of significance $\alpha = 0.05$)

Paired t-test

1. A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample.
2. Examples of where this might occur are:
   - Before-and-after observations on the same subjects (e.g. students' diagnostic test results before and after a particular module or course).
   - A comparison of two different methods of measurement or two different treatments where the measurements/treatments are applied to the same subjects (e.g. blood pressure measurements using a stethoscope and a dynamap).

Contd...



Sample
Before
After
Mean
Standard Deviation
Difference Zero?
Mean
Standard Deviation
Blood Pressure Before Treatment
Blood Pressure After Treatment

Assumptions for paired t-test

1. Two repeated or matched samples, In other words, must have a before and after design or matched pairs.
2. Paired samples T-tests can have only two groups.
3. Paired t-test assumes no extreme outliers.
4. The sampling distribution of the dependent variable should be normally distributed

What is the Hypothesis of Paired T-test?

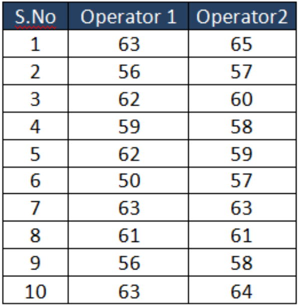
1. The Null Hypothesis for a paired t-test is that the average difference between the two population means is zero (0). In other words, there is no significant difference between the two population means.
2. The Alternative Hypothesis for a paired t-test – there is a significant difference between the two population means.

How to conduct a Paired T-test?

1. Establish the Null Hypothesis and Alternative Hypothesis
2. Determine the significance level
3. Calculate the difference between each observation in the two groups
4. Then, compute the mean difference ($\bar{x}_d$)
5. Calculate the standard deviation of differences (s) and then calculate the standard error, i.e. $\frac{s}{\sqrt{n}}$ (where n is the sample size)
6. Compute the t-Statistic,
$$t = \frac{\bar{x_d}}{\frac{s}{\sqrt{n}}} \tag{66}$$
7. Determine t critical value with $n - 1$ degrees of freedom
8. Finally, interpret the result. If the test statistic falls in the critical region, then reject the null hypothesis.

Example: Two operators are checking the same dimension on the same sample of 10 parts. Below are the results. Is there a significant operator measurement error? Test at the 5% significance level.

1. Consider the data as follows

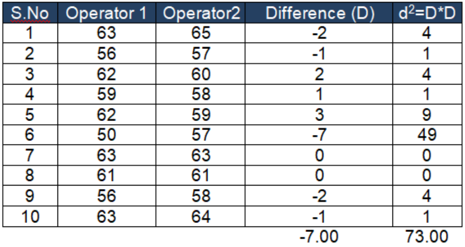


S.No	Operator 1	Operator2
1	63	65
2	56	57
3	62	60
4	59	58
5	62	59
6	50	57
7	63	63
8	61	61
9	56	58
10	63	64

2. We have,

Contd...

1. H_0 =There is no significant measurement error between the two operators.
   H_1 =There is a significant measurement error between the two operators.
2. $n = 10$, hence DOF=9
3. Create difference table and compute t-statistic



S.No	Operator 1	Operator2	Difference (D)	d^2=D*D
1	63	65	-2	4
2	56	57	-1	1
3	62	60	2	4
4	59	58	1	1
5	62	59	3	9
6	50	57	-7	49
7	63	63	0	0
8	61	61	0	0
9	56	58	-2	4
10	63	64	-1	1
			-7.00	73.00

$$t = \frac{\bar{x}_d}{\frac{s}{\sqrt{n}}} \tag{67}$$

4. Compute standard deviation of the differences (d) and compute standard error $SE = \frac{s}{\sqrt{n}}$

Contd...

1. We get $s = 2.7507$, and standard error $SE = \frac{2.7507}{\sqrt{10}}$
2. Hence, t-statistic is given as

$$t = \frac{-0.7}{0.8698} \quad (68)$$

$$= -0.8047 \quad (69)$$

3. Now, determine t critical value with $n - 1$ degrees of freedom
4. Since this is a Two-tailed test at α of 5% $t_{crit} = 2.262$
5. Compare t-statistic to t-critical $0.805 < 2.262$. In Hypothesis Testing, a critical value is a point on the test distribution that is compared to the test statistic to determine whether to reject the Null Hypothesis. The t calculated is not in the rejection region. Hence, we fail to reject the Null Hypothesis and say that there is no difference between the two mean values.

More examples

https://sixsigmastudyguide.com/paired-t-distribution-paired-t-test/

Example

1. An ambulance service company claims that on an average it takes 20 minutes between a call for ambulance and the patient's arrival at the hospital. If for 6 calls the time taken between call and arrival of patient in hospital are 27, 18, 26, 15, 20 and 32 minutes. Can the company's claim be accepted at 10% level of significance?
2. Ten students are selected from school and their heights are found to be 50, 52, 52, 53, 55, 56, 57, 58, 58 and 59 inches. In the light of this data, discuss the suggestion that the mean height of the school students is 54 inches at 5% level of significance.

Example

1. The specimen of copper wires drawn from a large lot have the following breaking strength (in kg weight) 578, 572, 570, 568, 572, 578, 570, 572, 596, 544. Test whether the men breaking strength of the lot may be taken as 578 kg. weight at 5% level of significance.
2. A certain stimulus administered to each of 12 patients resulted in the following increase of blood pressure: 5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6. Can it be concluded that the stimulus will in general be accompanied by an increase in blood pressure,

,

Contd...

- In one hour run of 16 test, the gasoline consumption of an engine averaged 16.4 gallons with standard deviation of 2.3 gallons. Test the claim that the average gasoline consumption of the engine is 12 gallons per hour.
- A machinist is making engine parts with axle diameter of 0.7 inches. A random sample of 10 parts shows the mean diameter 0.712 inch with standard deviation of 0.04 inch. On the basis of this sample, would you say that the work is inferior.

Chi-Square Test

1. The Chi-Square test is a statistical procedure for determining the difference between observed and expected data.
2. This test can also be used to determine whether it correlates to the categorical variables in our data.
3. It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.

Contd...

1. A chi-square test is a statistical test that is used to compare observed and expected results.
2. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration.
3. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.

Contd...

1. A chi-square test or comparable non-parametric test is required to test a hypothesis regarding the distribution of a categorical variable.
2. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal.
3. They cannot have a normal distribution since they can only have a few particular values.
4. For example, a meal delivery firm in India wants to investigate the link between gender, geography, and people's food preferences.

Formula for Chi-Square Test

1. Given the expected and observed frequency, χ^2 statistic is given as

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{70}$$

2. where c is degree of freedom, O_i is the observed frequency, and E_i is the expected value
3. The Observed values are those you gather yourselves. The expected values are the frequencies expected, based on the null hypothesis.

Uses of Chi-Square Test

1. The Chi-squared test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution.
2. The Chi-squared test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets.

Types of Chi-Square Test

1. Test for Independence: The Chi-Square Test of Independence is a derivable (also known as inferential) statistical test which examines whether the two sets of variables are likely to be related with each other or not.
2. This test is used when we have counts of values for two nominal or categorical variables and is considered as non-parametric test.
3. A relatively large sample size and independence of observations are the required criteria for conducting this test.

Contd...

1. In statistical hypothesis testing, the Chi-Square Goodness-of-Fit test determines whether a variable is likely to come from a given distribution or not.
2. We must have a set of data values and the idea of the distribution of this data. We can use this test when we have value counts for categorical variables.
3. This test demonstrates a way of deciding if the data values have a “good enough” fit for our idea or if it is a representative sample data of the entire population.

Example

1. Suppose you want to determine whether there is a significant association between gender and smoking status.
2. You randomly sample 200 individuals and obtain the following data

	Smoker	Non-Smoker	Total
Female	20	80	100
Male	50	50	100
Total	70	130	200

3. Next, you can calculate the expected frequencies under the null hypothesis of no association between gender and smoking status. The expected frequency for each cell is equal to (row total x column total) / sample size.

Contd...

1. Expected frequencies can be given as

	Smoker	Non-Smoker	Total
Female	35	65	100
Male	35	65	100
Total	70	130	200

2. Compute χ^2-statistic using the formula

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{71}$$

$$= \frac{(20-35)^2}{35} + \frac{(80-65)^2}{65} + \frac{(50-35)^2}{35} + \frac{(50-65)^2}{65} \tag{72}$$

$$= 7.21 \tag{73}$$

3. The degrees of freedom for this test are equal to (number of rows - 1) x (number of columns - 1) = 1 x 1 = 1. Using a chi-square distribution table or a calculator, we can find the critical value for a significance level of 0.05 and 1 degree of freedom to be 3.84.

Chi-Square Test for Homogeneity

1. The chi-square test of homogeneity is the non-parametric test used in a situation where the dependent variable is categorical. Data can be presented using a contingency table in which populations and categories of the variable are the row and column labels.
2. The “test of homogeneity” is a way of determining whether two or more sub-groups of a population share the same distribution of a single categorical variable.
3. For example, do people of different races have the same proportion of smokers to non-smokers, or do different education levels have different proportions of Democrats, Republicans, and Independent. The homogeneity test is used if the response variable has several outcome categories, and we wish to compare two or more groups.

How to check test for Homogeneity and Independence?

1. The difference is a matter of design. In the test of independence, observational units are collected at random from a population and two categorical variables are observed for each unit.
2. In the test of homogeneity, the data are collected by randomly sampling from each sub-group separately.

Example

1. At a major credit card bank, the percentages of people who historically apply for the Silver, Gold, and Platinum cards are 60%, 30%, and 10%, respectively. In a recent sample of customers responding to a promotion, of 200 customers, 110 applied for Silver, 55 for Gold, and 35 for Platinum. Is there evidence to suggest that the percentages for this promotion may be different from the historical proportions?
2. What is the expected number of customers applying for each type of card in this sample if the historical proportions are still true?
3. Compute the χ^2 statistic.
4. How many degrees of freedom does the χ^2 statistic have?

Example

1. Human births If there is no seasonal effect on human births, we would expect equal numbers of children to be born in each season (winter, spring, summer, and fall). A student takes a census of her statistics class and finds that of the 120 students in the class, 25 were born in winter, 35 in spring, 32 in summer, and 28 in fall. She wonders if the excess in the spring is an indication that births are not uniform throughout the year.
2. What is the expected number of births in each season if there is no "seasonal effect" on births?
3. Compute the χ^2 statistic.
4. How many degrees of freedom does the χ^2 statistic have?

Example

1. Car origins: A random survey of autos parked in the student lot and the staff lot at a large university classified the brands by country of origin, as seen in the table. Are there differences in the national origins of cars driven by students and staff?

	Student	Staff
American	107	105
European	33	12
Asian	55	47

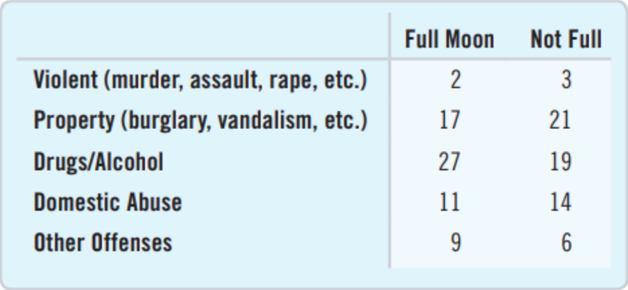
2. Is this a test of independence or homogeneity?
3. Write appropriate hypotheses.

Example

1. Mileage A salesman who is on the road visiting clients thinks that, on average, he drives the same distance each day of the week. He keeps track of his mileage for several weeks and discovers that he averages 132 miles on Mondays, 213 miles on Tuesdays, 168 miles on Wednesdays, 180 miles on Thursdays, and 103 miles on Fridays. He wonders if this evidence contradicts his belief in a uniform distribution of miles across the days of the week.
2. Explain why it is not appropriate to test his hypothesis using the chi-square goodness-of-fit test.

Example

1. Full moon: Some people believe that a full moon elicits unusual behavior in people. The table shows the number of arrests made in a small town during weeks of six full moons and six other randomly selected weeks in the same year. We wonder if there is evidence of a difference in the types of illegal activity that take place



	Full Moon	Not Full
Violent (murder, assault, rape, etc.)	2	3
Property (burglary, vandalism, etc.)	17	21
Drugs/Alcohol	27	19
Domestic Abuse	11	14
Other Offenses	9	6

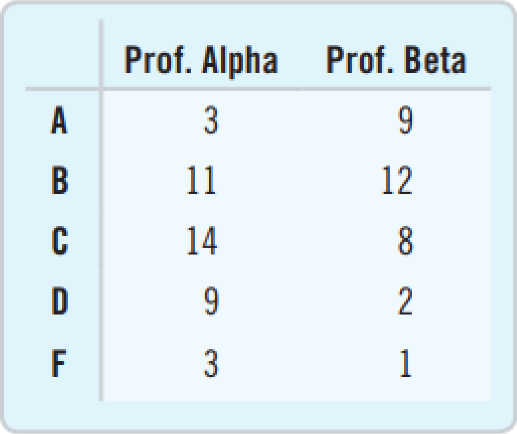
2.

Contd...

1. Will you test goodness-of-fit, homogeneity, or independence?
2. Write appropriate hypotheses.
3. Find the expected counts for each cell, and explain why the chi-square procedures are not appropriate

Example

1. Grades Two different professors teach an introductory Statistics course. The table shows the distribution of final grades they reported. We wonder whether one of these professors is an “easier” grader.



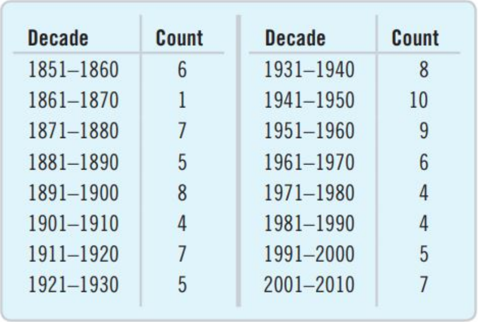
	Prof. Alpha	Prof. Beta
A	3	9
B	11	12
C	14	8
D	9	2
F	3	1

Contd...

1. Will you test goodness-of-fit, homogeneity, or independence?
2. Write appropriate hypotheses.
3. Find the expected counts for each cell, and explain why the chi-square procedures are not appropriate.

Example

1. Hurricane frequencies: The National Hurricane Center provides data that list the numbers of large (category 3, 4, or 5) hurricanes that have struck the United States, by decade since 1851 (www.nhc.noaa.gov/dcmi.shtml). The data are given below.



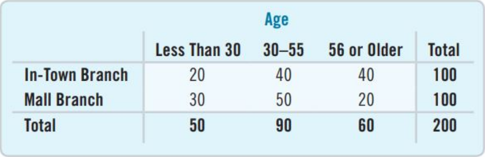
Decade	Count	Decade	Count
1851–1860	6	1931–1940	8
1861–1870	1	1941–1950	10
1871–1880	7	1951–1960	9
1881–1890	5	1961–1970	6
1891–1900	8	1971–1980	4
1901–1910	4	1981–1990	4
1911–1920	7	1991–2000	5
1921–1930	5	2001–2010	7

Contd...

1. Recently, there's been some concern that perhaps the number of large hurricanes has been increasing. The natural null hypothesis would be that the frequency of such hurricanes has remained constant.
   a) With 96 large hurricanes observed over the 16 periods, what are the expected value(s) for each cell?
2. What kind of chi-square test would be appropriate?
3. State the null and alternative hypotheses.
4. How many degrees of freedom are there?
5. The value of χ^2 is 12.67. What's the P-value?
6. State your conclusion.

Example

1. Customer ages An analyst at a local bank wonders if the age distribution of customers coming for service at his branch in town is the same as at the branch located near the mall. He selects 100 transactions at random from each branch and researches the age information for the associated customer. Here are the data:



	Age			
	Less Than 30	30–55	56 or Older	Total
In-Town Branch	20	40	40	**100**
Mall Branch	30	50	20	**100**
Total	**50**	**90**	**60**	**200**

Contd...

1. What is the null hypothesis?
2. What type of test is this?
3. What are the expected numbers for each cell if the null hypothesis is true?
4. Find the χ^2 statistic.
5. How many degrees of freedom does it have?
6. Find the P-value.
7. What do you conclude?