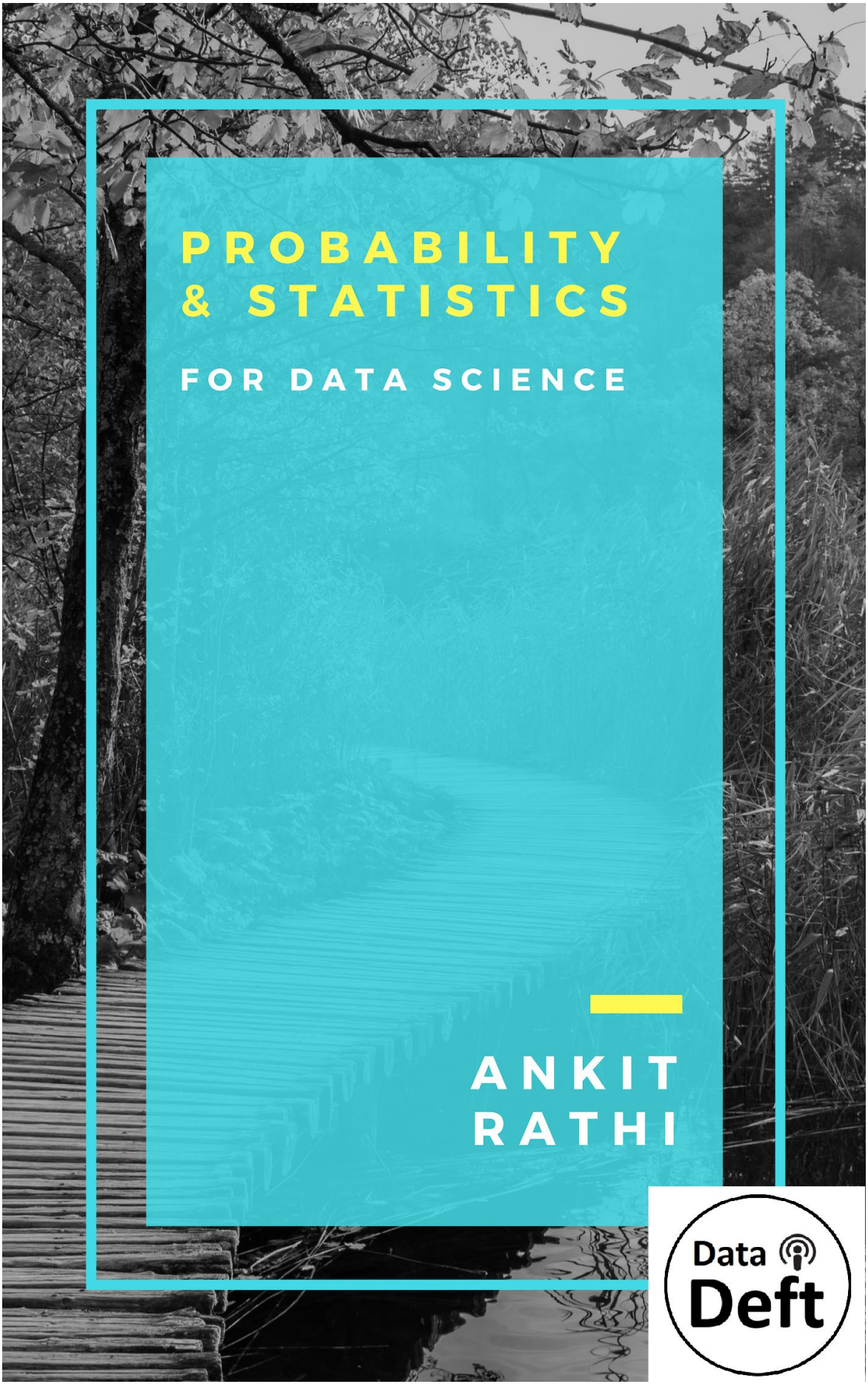

PROBABILITY
& STATISTICS
FOR DATA SCIENCE
ANKIT
RATHI
Data
Deft

Probability & Statistics

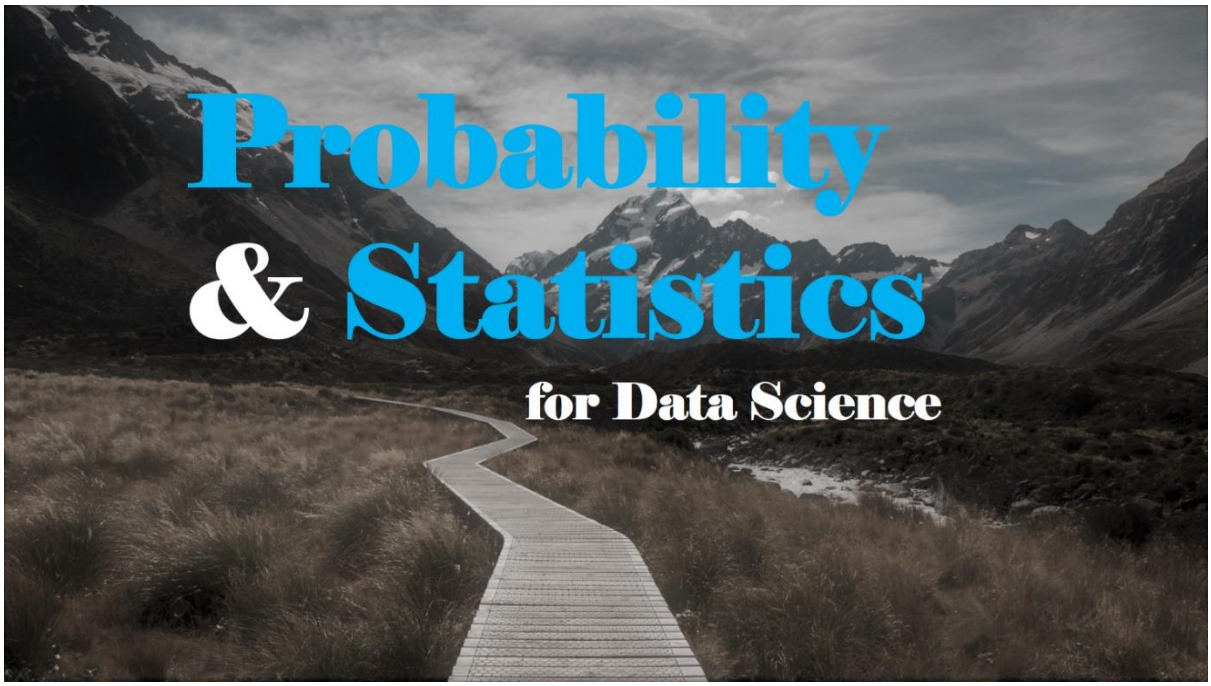
for Data Science

Ankit Rathi

Table of Contents

Chapter 1

Introduction



This is the first chapter which covers the context and topics covered in this book.

- *Probability*
- *Descriptive Statistics*
- *Inferential Statistics*
- *Bayesian Statistics*
- *Statistical Learning*

I haven't attended any formal education in probability & statistics, whatever I have learnt in bits and pieces till now is through working on data science problems. When I look at the literature available on probability & statistics, I find it too theoretical and generalized. I have felt that there should be some literature on probability & statistics specifically focused on data science.

Recently couple of books have been written like '*Practical Statistics for Data Scientists: 50 Essential Concepts*' by *Peter Bruce/ Andrew Bruce*, which are good and cover some of the context, but I want to cover everything about probability & statistics from basics to statistical learning. I would like to mention that my focus in these posts would be to give intuition on every topic and how it relates to data science rather going deep into mathematical formulas.

This book contains 6 chapters, this one is the first which gives an overview and set the context of subsequent chapters.

Second chapter describes probability & its types, random variables & probability distributions and how they are important from data science perspective.

Probability

- Introduction
- Conditional Probability
- Random Variables
- Probability Distributions

Third, Fourth & Fifth chapters cover every topic related to statistics & its significance in data science.

Statistics

- Introduction
- Descriptive Statistics
- Inferential Statistics
- Bayesian Statistics

Sixth (& final) chapter elaborates statistical learning, it will be about looking at machine learning or data science from statistical perspective.

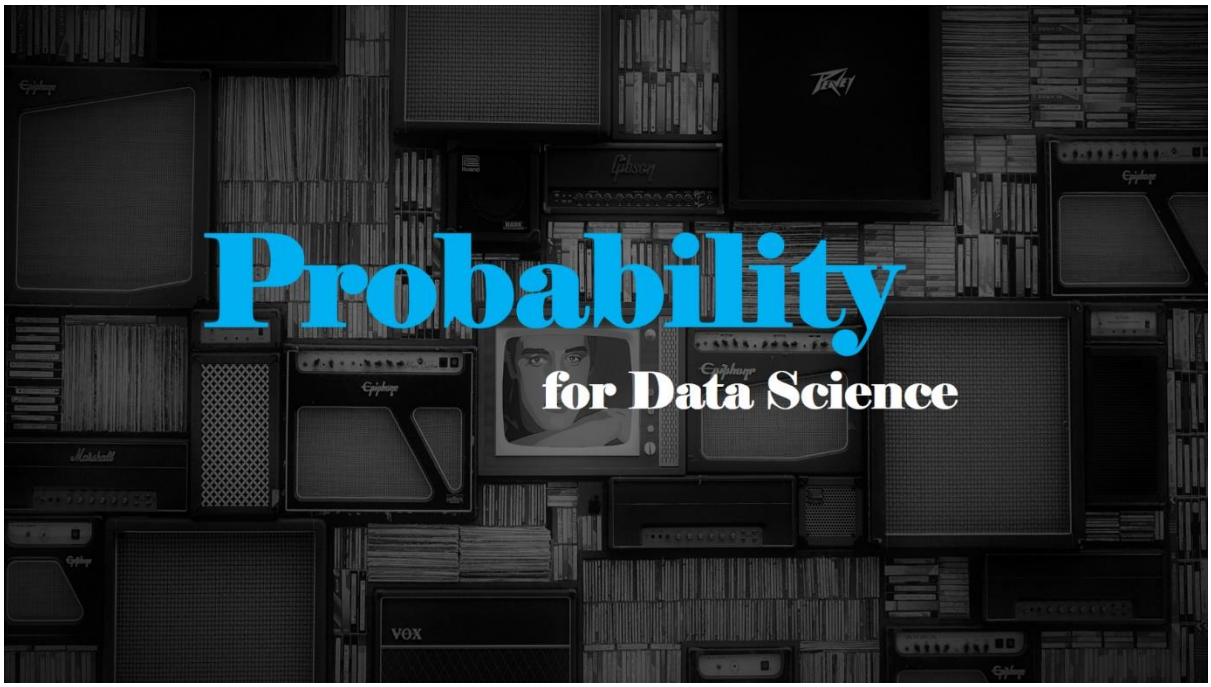
Statistical Learning

- Introduction
- Prediction & Inference
- Parametric & Non-parametric methods
- Prediction Accuracy and Model Interpretability
- Bias-Variance Trade-Off

So if you are looking for similar kind of learning curve, kindly continue with subsequent chapters.

Chapter 2

Probability

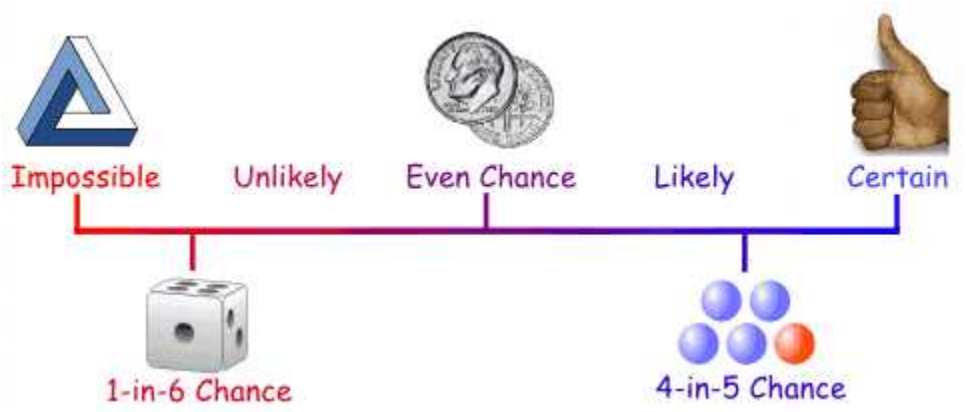


This chapter covers these topics related to probability and their significance in data science.

- *Introduction*
- *Conditional Probability*
- *Random Variables*
- *Probability Distributions*

What is probability?

Probability is the chance that something will happen—how likely it is that some event will happen.



Probability of an event happening ***P(E)*** *= Number of ways it can happen* ***n(E)****/ Total number of outcomes* ***n(T)***

Probability is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1, where 0 indicates impossibility and 1 indicates certainty.

Probability
Math explained in easy language, plus puzzles, games, quizzes, worksheets and a forum. For K-12 kids, teachers and...
www.mathsisfun.com

Why probability is important?

Uncertainty and randomness occur in many aspects of our daily life and having a good knowledge of probability helps us make sense of these uncertainties. Learning about probability helps us make informed judgments on what is likely to happen, based on a pattern of data collected previously or an estimate.

Why is Learning About Probability Important?
Uncertainty and randomness occur in many aspects of our daily life and having a good knowledge of probability helps us...
medium.com

How Probability is used in Data Science?

Data science often uses statistical inferences to predict or analyse trends from data, while statistical inferences uses probability distributions of data. Hence knowing probability and its applications are important to work effectively on data science problems.

What is the use of probability in data science?
Answer (1 of 2): Probability is the foundation and language needed for most of statistics. Understanding the methods...
www.quora.com

How important knowing Probability and Statistics is for Data Science?
Data Science is a subject that is dear to me and I have found a lot of resonance, since I quit my corporate outsourcing...
www.linkedin.com

What is Conditional Probability?

Conditional probability is a measure of the probability of an event (some particular situation occurring) given that (by assumption, presumption, assertion or evidence) another event has occurred.

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

The probability of event B given event A equals the probability of event A and event B divided by the probability of event A.

Conditional Probability
Math explained in easy language, plus puzzles, games, quizzes, worksheets and a forum. For K-12 kids, teachers and...
www.mathsisfun.com

Introduction to Conditional Probability and Bayes theorem for data science professionals
Introduction Understanding of probability is must for a data science professional. Solutions to many data science...
www.analyticsvidhya.com

How conditional probability is used in data science?

Many data science techniques (i.e. Naive Bayes) rely on Bayes' theorem. Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence.

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

Using the Bayes' theorem, its possible to build a learner that predicts the probability of the response variable belonging to some class, given a new set of attributes.

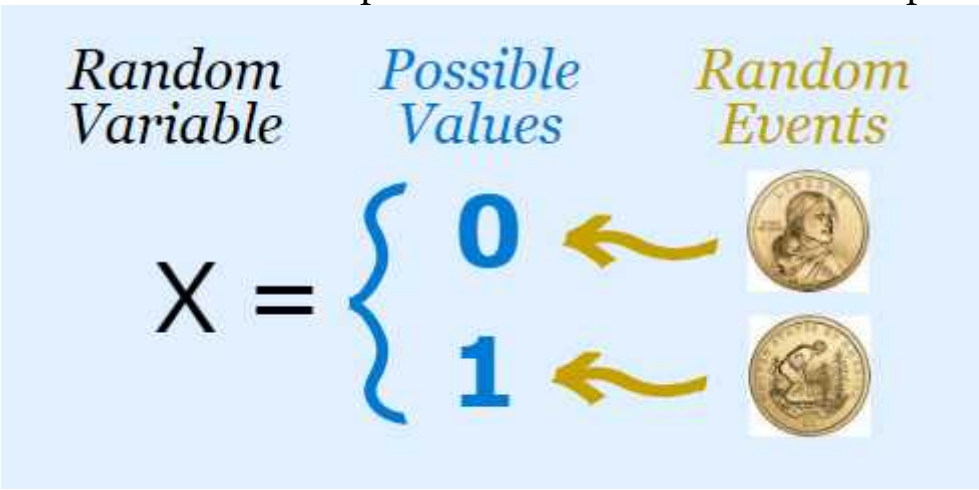
$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Bayes' Theorem and Conditional Probability | Brilliant Math & Science Wiki
Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence. It...
brilliant.org

Bayes' Rule Applied
Using Bayesian Inference on a real-world problem
towardsdatascience.com

What are random variables?

A random variable is a set of possible values from a random experiment.



A random variable (random quantity, aleatory variable, or stochastic variable) is a variable whose possible values are outcomes of a random phenomenon.

Random variables can be discrete or continuous. Discrete random variables can only take certain values while continuous random variables can take any value (within a range).

Random Variables
Math explained in easy language, plus puzzles, games, quizzes, worksheets and a forum. For K-12 kids, teachers and...
www.mathsisfun.com

Random Variables
random variable , usually written X, is a variable whose possible values are numerical outcomes of a random phenomenon...
www.stat.yale.edu

What are probability distributions?

The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable.

For a discrete random variable, x, the probability distribution is defined by a probability mass function, denoted by f(x). This function provides the probability for each value of the random variable.

For a continuous random variable, since there is an infinite number of values in any interval, the probability that a continuous random variable will lie within a given interval is considered. So here, the probability distribution is defined by probability density function, also denoted by f(x).

Both probability functions must satisfy two requirements:

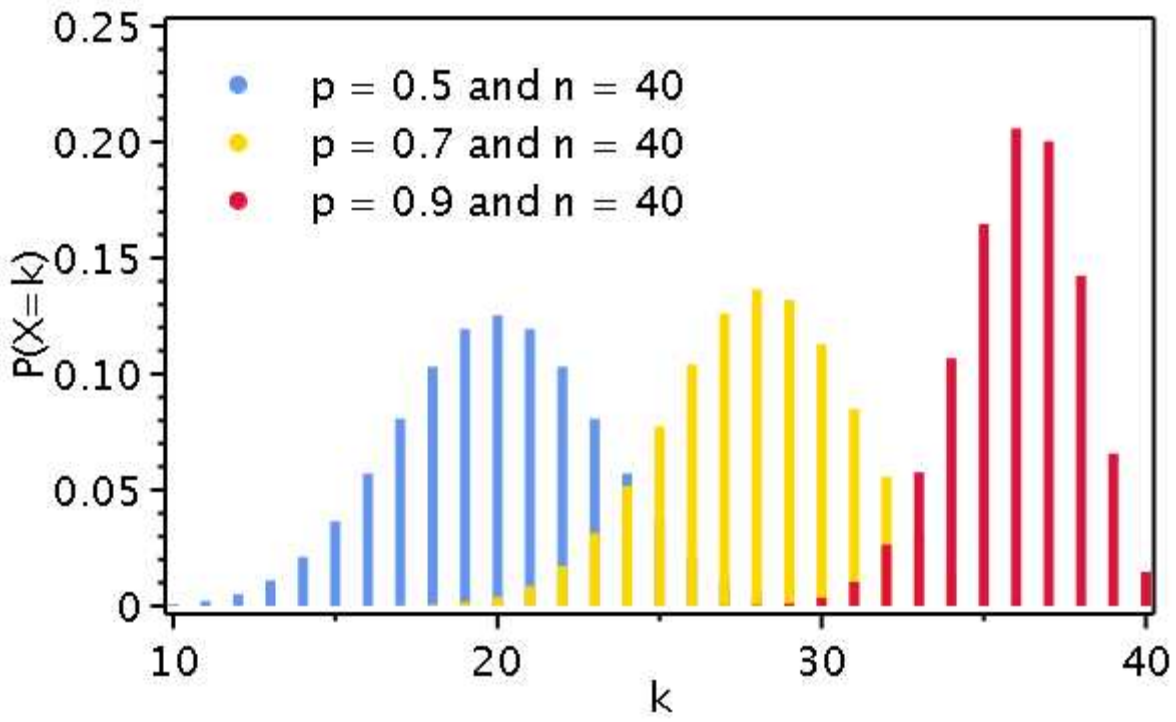
(1) $f(x)$ must be non-negative for each value of the random variable

(2) the sum of the probabilities for each value (or integral over all values) of the random variable must equal one.

Statistics — Random variables and probability distributions
Statistics — Random variables and probability distributions: A random variable is a numerical description of the...
www.britannica.com

What are the types of probability distributions?

A binomial distribution is a statistical experiment that has the following properties: The experiment consists of n repeated trials. Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure. The probability of success, denoted by P, is the same on every trial.



Binomial Distribution

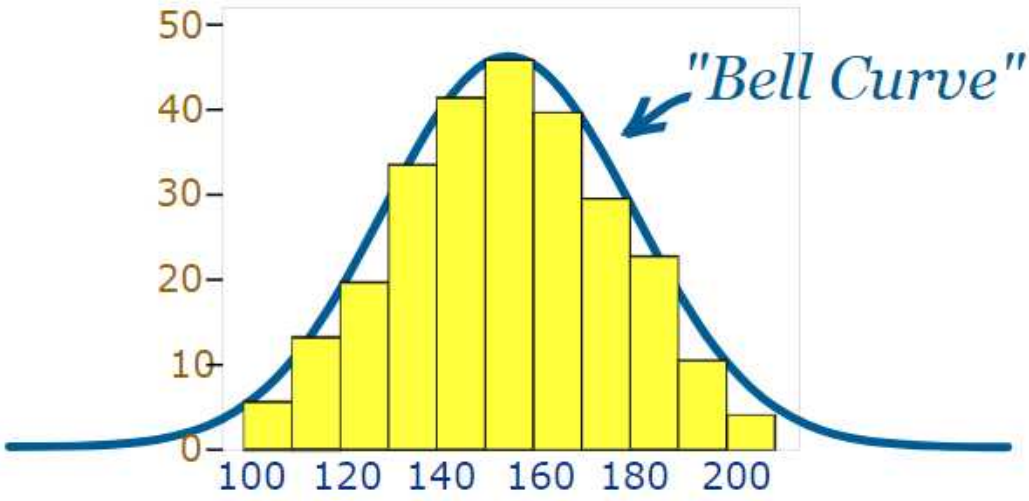
Probability of k out of n ways:

$$P(k \text{ out of } n) = \frac{n!}{k!(n-k)!} p^k(1-p)^{(n-k)}$$

The General Binomial Probability Formula

The normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. It has following properties:

- The normal curve is symmetrical about the mean μ;
- The mean is at the middle and divides the area into halves;
- The total area under the curve is equal to 1;
- It is completely determined by its mean and standard deviation σ (or variance σ2)



A Normal Distribution

For other common *probability distributions*, please refer following links:

6 Common Probability Distributions every data science professional should know
Introduction Suppose you are a teacher in a university. After checking assignments for a week, you graded all the…
www.analyticsvidhya.com

Probability Distribution: List of Statistical Distributions
Statistics Definitions > A probability distribution tells you what the probability of an event happening is…
www.statisticshowto.com

How random variables & probability distributions are used in data science?

Data science often uses statistical inferences to predict or analyse trends from data, while statistical inferences uses probability distributions of data. Hence knowing random variables & their probability distributions are important to work effectively on data science problems.

How to Dominate the Statistics Portion of Your Data Science Interview
For someone working or trying to work in data science, statistics is probably the biggest and most intimidating area of...
www.datascience.com

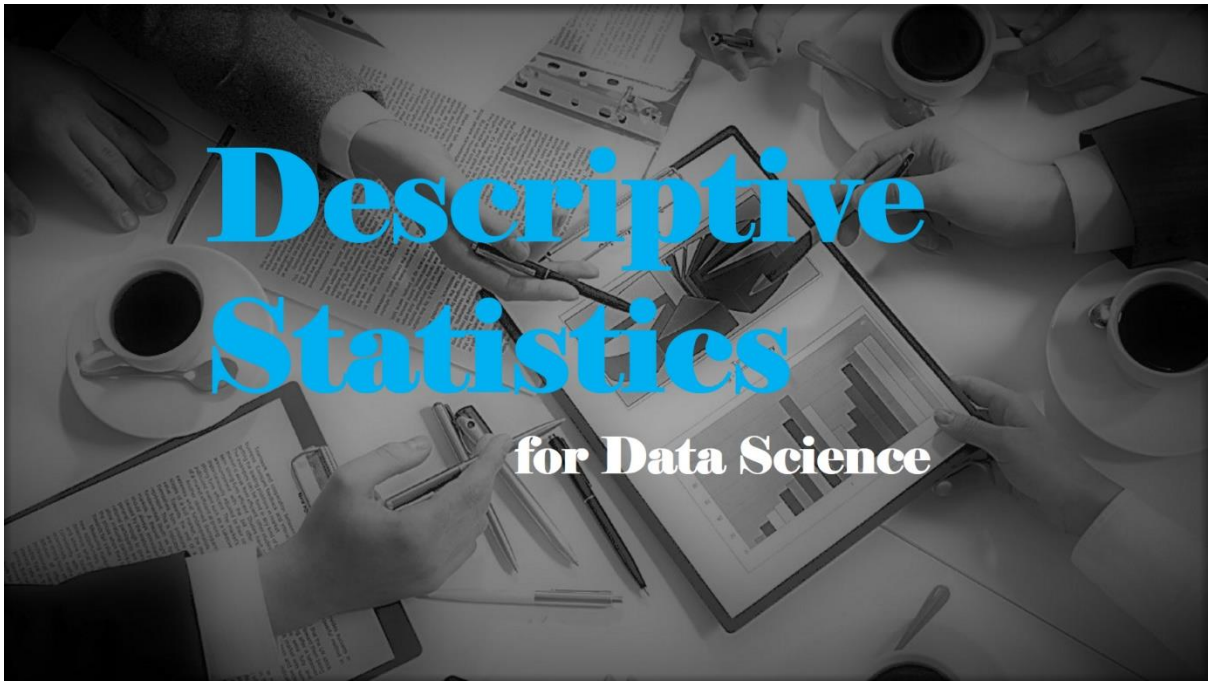
Basics of Probability for Data Science explained with examples
Statistically, the probability of any one of us being here is so small that you'd think the mere fact of existing would...
www.analyticsvidhya.com

What are probability distributions used for?
Answer (1 of 5): Well, probability is the basis for most of statistics, and statistics is mostly a real-world pursuit...
www.quora.com

Most ML models assume an underlying data distribution for them to function well. Where can I learn...
Answer (1 of 9): The question is: >> Most ML models assume an underlying data distribution for them to function well...
www.quora.com

Chapter 3

Descriptive Statistics



This chapter covers these topics related to descriptive statistics and their significance in data science.

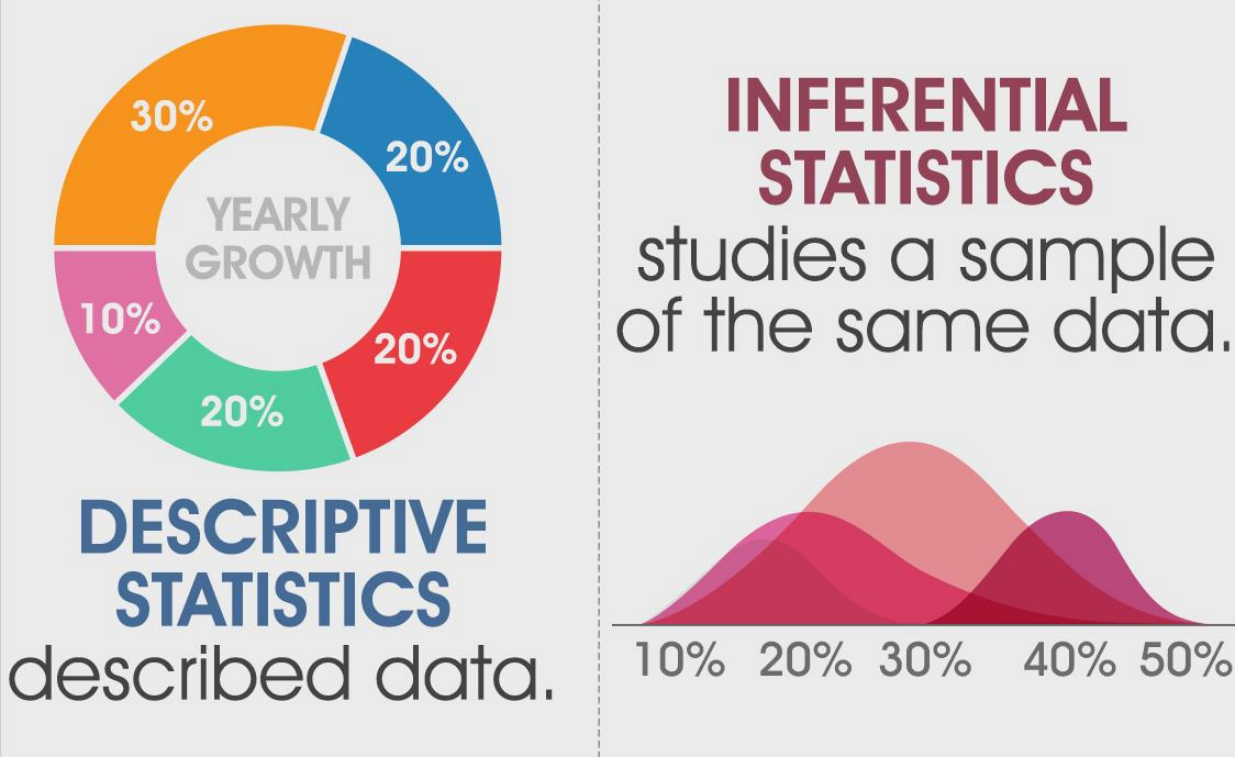
- *Introduction to Statistics*
- *Descriptive Statistics*
- *Uni-variate Analysis*
- *Bi-variate Analysis*
- *Multivariate Analysis*
- *Function Models*
- *Significance in Data Science*

Statistics Introduction

Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.

Informal definition :D

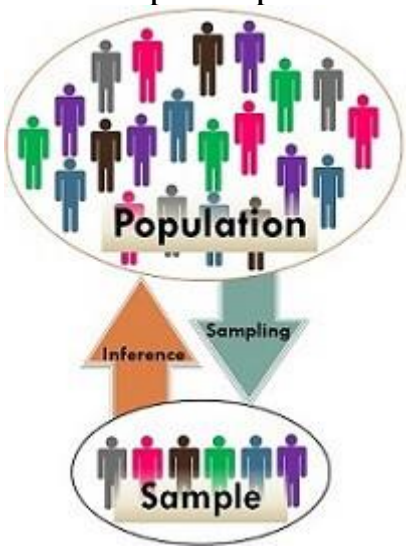
Statistics, in short, is the study of data. It includes descriptive statistics (the study of methods and tools for collecting data, and mathematical models to describe and interpret data) and inferential statistics (the systems and techniques for making probability-based decisions and accurate predictions.



Descriptive Vs Inferential Statistics

Population vs Sample

Population means the aggregate of all elements under study having one or more common characteristic while sample is a part of population chosen at random for participation in the study.



Population Vs Sample

Descriptive Statistics

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information. Descriptive statistics are just descriptive. They do not involve generalizing beyond the data at hand.

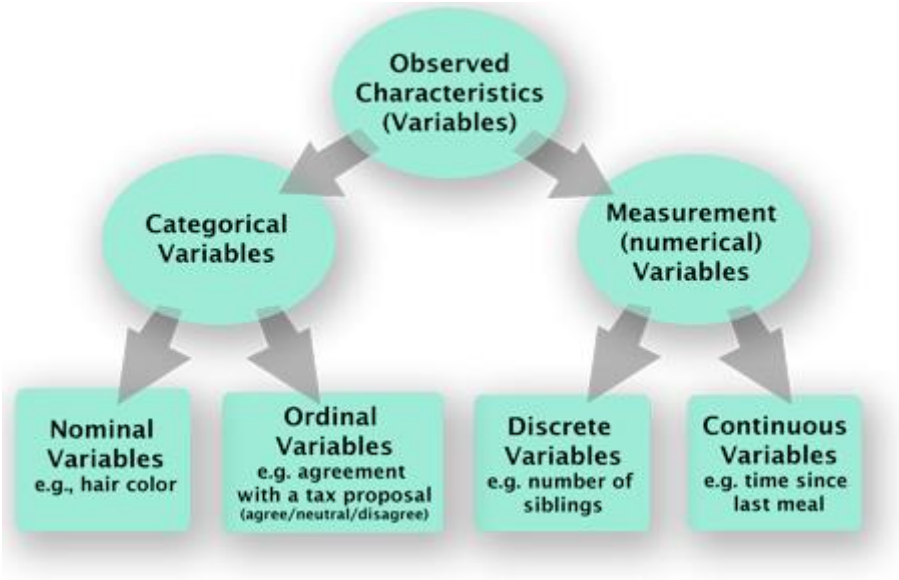
Types of Variable

Dependent and Independent Variables: An independent variable (experimental or predictor) is a variable that is being manipulated in an experiment in order to observe the effect on a dependent variable (outcome).

Dependent Variable	Independent Variable
Explained	Explanatory
Predictand	Predictor
Regressand	Regressor
Response	Stimulus
Outcome	Covariate
Controlled	Control

Other names of variables

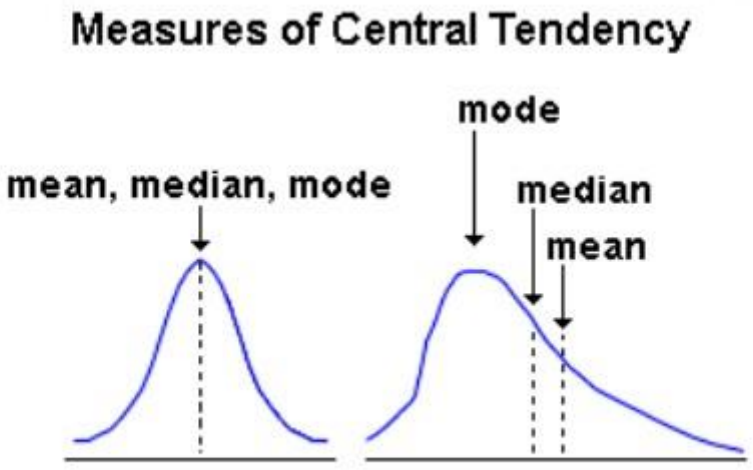
Categorical and Continuous Variables: Categorical variables (qualitative) represent types of data which may be divided into groups. Categorical variables can be further categorized as either nominal, ordinal or dichotomous. Continuous variables (quantitative) can take any value. Continuous variables can be further categorized as either interval or ratio variables.



Categorical Vs continuous variables

Central Tendency

Central tendency is a central or typical value for a distribution. It may also be called a centre or location of the distribution. The most common measures of central tendency are the arithmetic mean, the median and the mode.

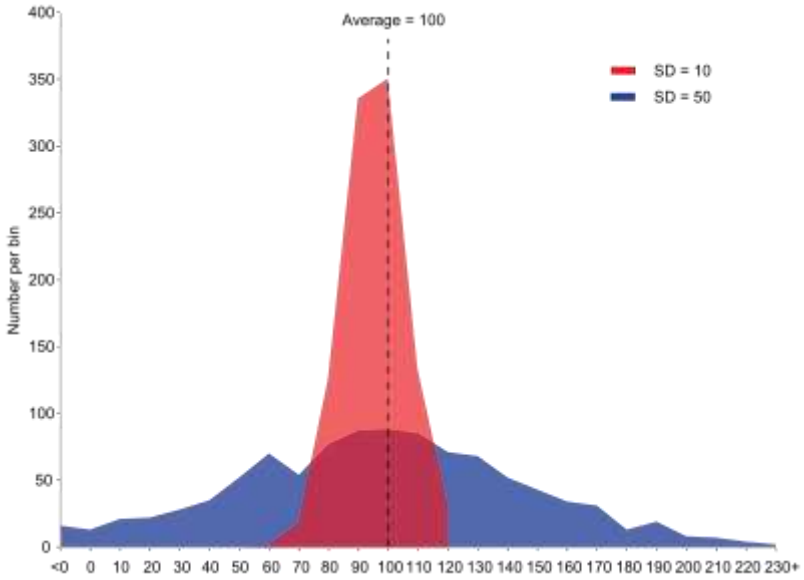


Mean, median & mode as Central Tendency

Mean is the numerical average of all values, median is directly in the middle of the data set while mode is the most frequent value in the data set.

Spread or Variance

Spread (dispersion or variability) is the extent to which a distribution is stretched or squeezed. Common examples of measures of statistical dispersion are the variance, standard deviation, and inter-quartile range (IQR).

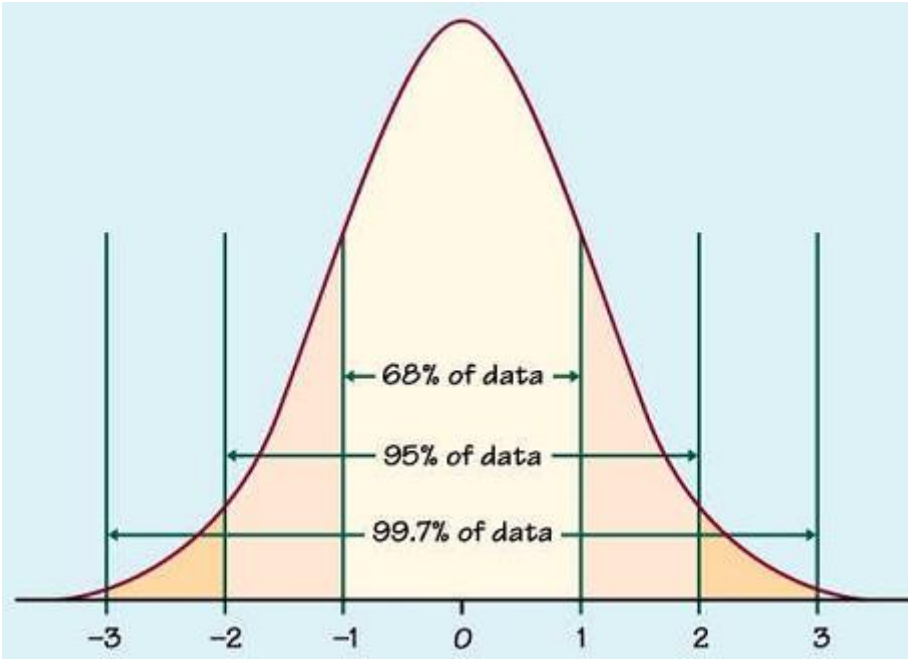


Spread or Variance

Inter-quartile range (IQR) is the distance between the 1st quartile and 3rd quartile and gives us the range of the middle 50% of our data. Variance is the average of the squared differences from the mean while standard deviation is the square root of the variance.

Upper outliers: Q3+1.5 ·IQR
Lower outliers: Q1–1.5 ·IQR

Standard Score or Z score: For an observed value x, the Z score finds the number of standard deviations x is away from the mean.



Standard deviation & Z-score

The Central Limit Theorem is used to help us understand the following facts regardless of whether the population distribution is normal or not:

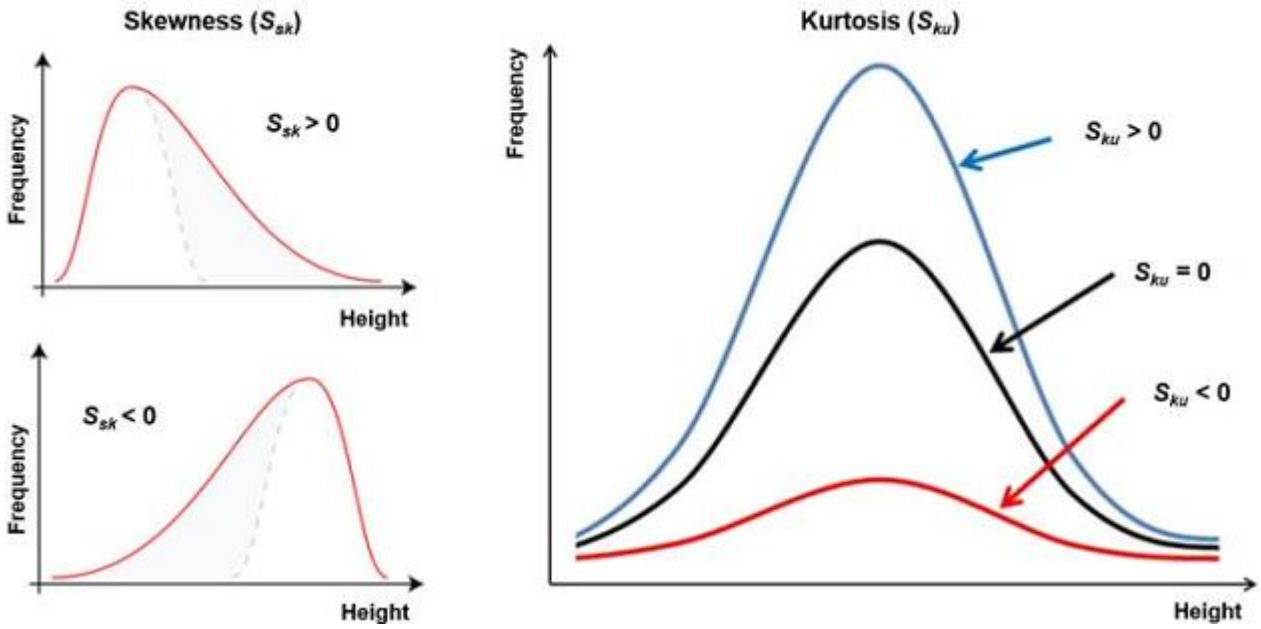
1. the mean of the sample means is the same as the population mean
2. the standard deviation of the sample means is always equal to the standard error.
3. the distribution of sample means will become increasingly more normal as the sample size increases.

Univariate Analysis

In univariate analysis, appropriate statistic depends on the level of measurement. For nominal variables, a frequency table and a listing of the mode(s) is sufficient. For ordinal variables the median can be calculated as a measure of central tendency and the range (and variations of it) as a measure of dispersion. For interval level variables, the arithmetic mean (average) and standard deviation are added to the toolbox and, for ratio level variables, we add the geometric mean and harmonic mean as measures of central tendency and the coefficient of variation as a measure of dispersion.

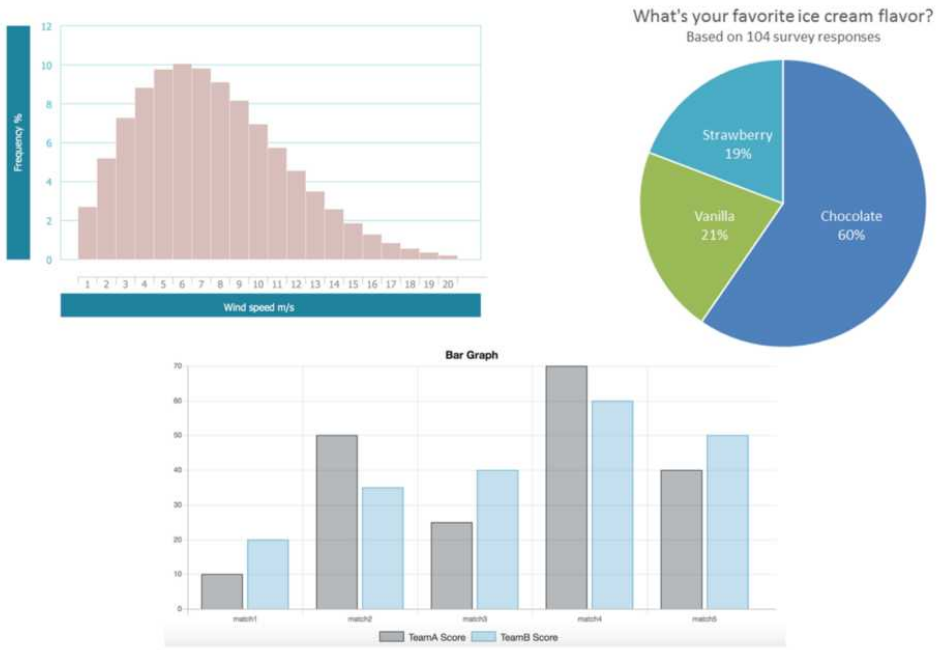
For interval and ratio level data, further descriptors include the variable's skewness and kurtosis. Skewness is a measure of symmetry, or more

precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.



Skewness & Kurtosis

Mainly, bar graphs, pie charts and histograms are used for uni-variate analysis.

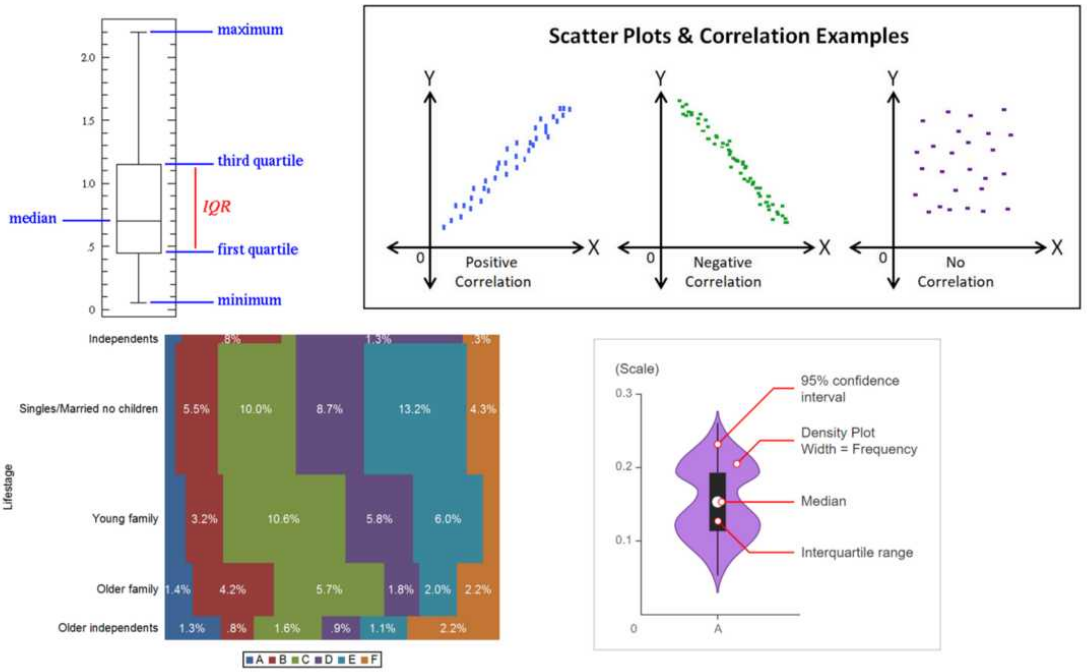


Histogram, Pie-chart & Bar-graph

Bi-variate Distribution

Bivariate analysis involves the analysis of two variables (often denoted as X, Y), for determining the empirical relationship between them.

For two continuous variables, a scatter-plot is a common graph. When one variable is categorical and the other continuous, a box-plot or violin-plot (also Z-test and t-test) is common and when both are categorical a mosaic plot is common (also chi-square test).



Box-plot, Scatter-plot, Mosaic-plot & Violin-plot

z-Test

Formula to find the value of Z (z-test) Is:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$\bar{x}$ = mean of sample

μ_0 = mean of population

σ = standard deviation of population

n = no. of observations

t-Test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \begin{matrix} \mapsto \\ \mapsto \\ \mapsto \end{matrix} \quad \frac{\text{difference between means}}{\frac{\text{variance}}{\text{sample size}}}$$

where $\bar{x}_1$ = mean of sample 1

$\bar{x}_2$ = mean of sample 2

n_1 = number of subjects in sample 1

n_2 = number of subjects in sample 2

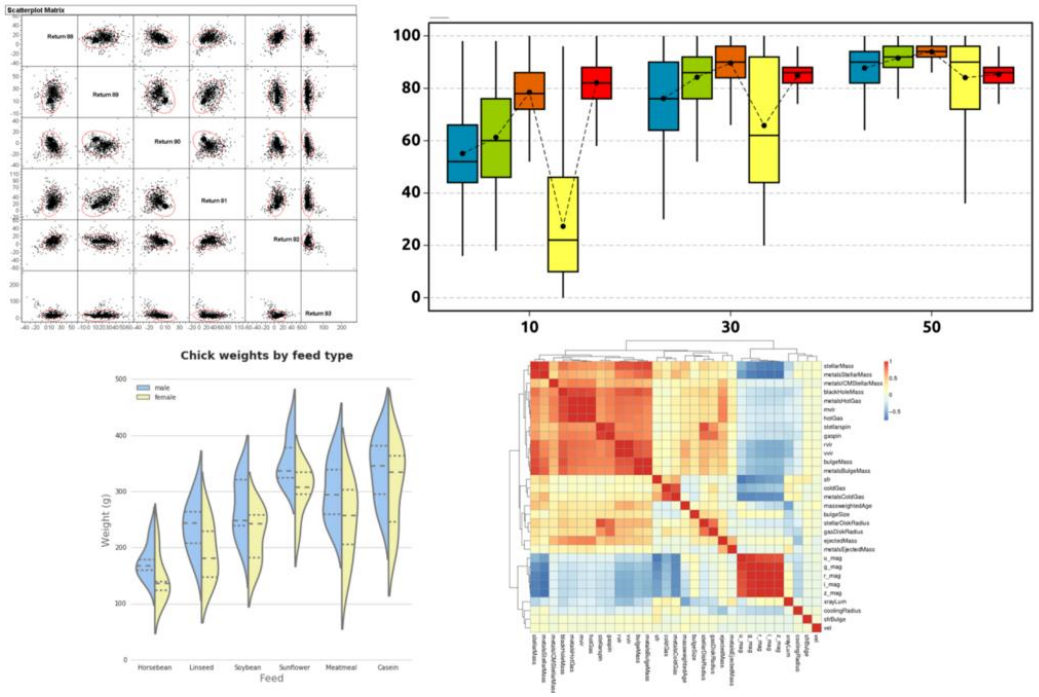
s_1^2 = variance of sample 1 = $\frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1}$

s_2^2 = variance of sample 2 = $\frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2}$

z & t-Tests

Multivariate Analysis

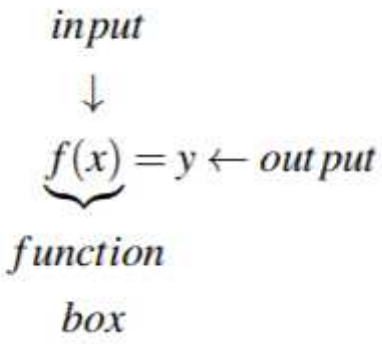
Multivariate analysis involves observation and analysis of more than one statistical outcome variable at a time. Multivariate scatter plot grouped box-plot (or grouped violin-plot), heat-map are used for multi-variate analysis.



Multi-variate scatter-plot, Grouped box-plot, Grouped violin-plot, Heat-map

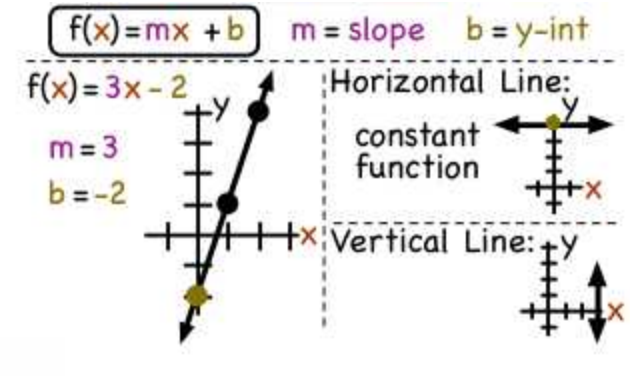
Function Models

A function can be expressed as an equation, as shown below. In the equation, *f* represents the function name and *x* represents the independent variable and *y* represents the dependent variable.



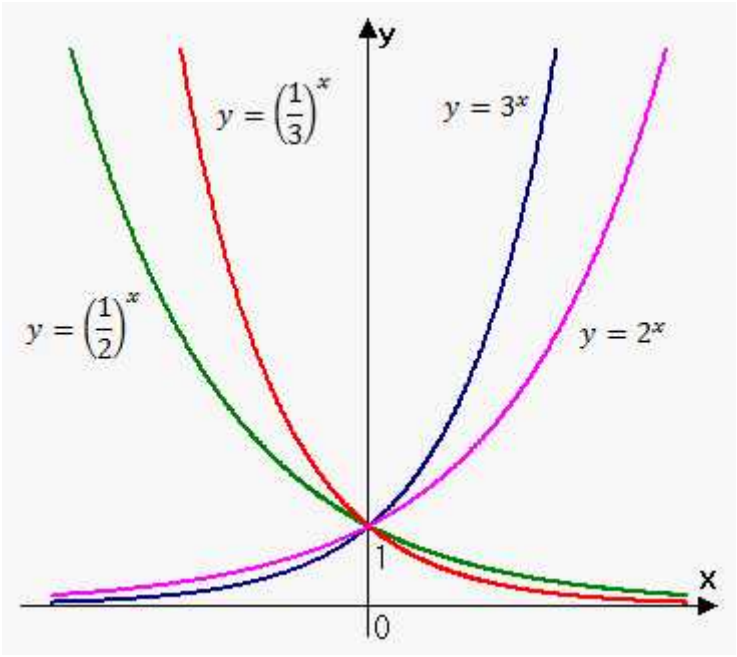
A typical function

A linear function has the same average rate of change on every interval. When a linear model is used to describe data, it assumes a constant rate of change.



Linear function

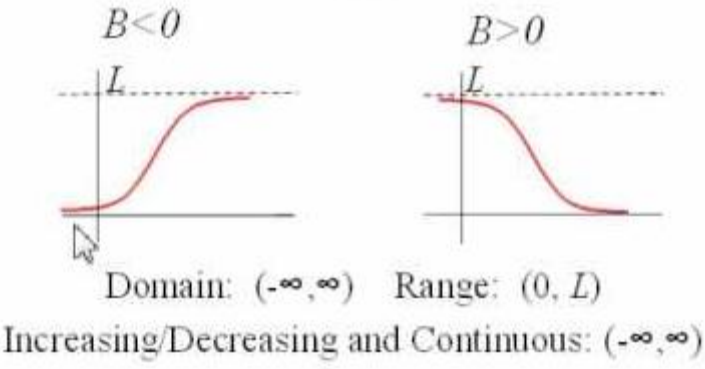
Exponential functions have variable appears as the exponent (or power) instead of the base.



Exponential functions

The logistic function has effect of limiting upper bound, a curve that grows exponentially at first and then slows down and hardly grows at all.

$$f(x) = \frac{L}{1 + Ae^{Bx}}$$



Logistic functions

Significance in Data Science

Descriptive Statistics helps you to understand your data and is initial & very important step of Data Science. This is since Data Science is all about making predictions and you can't predict if you can't understand the patterns in existing data.

References:

Udacity
Descriptive Statistics
classroom.udacity.com

Edx
Descriptive Statistics
courses.edx.org

Chapter 4

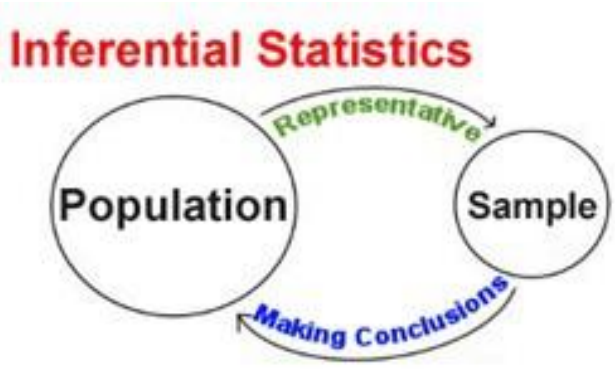
Inferential Statistics

This chapter covers these topics related to inferential statistics and their significance in data science.

- *Inferential Statistics*
- *Sampling Distributions & Estimation*
- *Hypothesis Testing (One and Two Group Means)*
- *Hypothesis Testing (Categorical Data)*
- *Hypothesis Testing (More Than Two Group Means)*
- *Quantitative Data (Correlation & Regression)*
- *Significance in Data Science*

Inferential Statistics

Inferential statistics allows you to make inferences about the population from the sample data.



Population & Sample

A sample is a representative subset of a population. Conducting a census on population is an ideal but impractical approach in most of the cases. Sampling is much more practical; however, it is prone to sampling error. A sample non-representative of population is called bias, method chosen for such sampling is called sampling bias. Convenience bias, judgement bias, size bias, response bias are main types of sampling bias. The best technique for reducing bias in sampling is randomization. Simple random sampling is the simplest of randomization techniques, cluster sampling & stratified sampling are other systematic sampling techniques.

Sampling Distributions

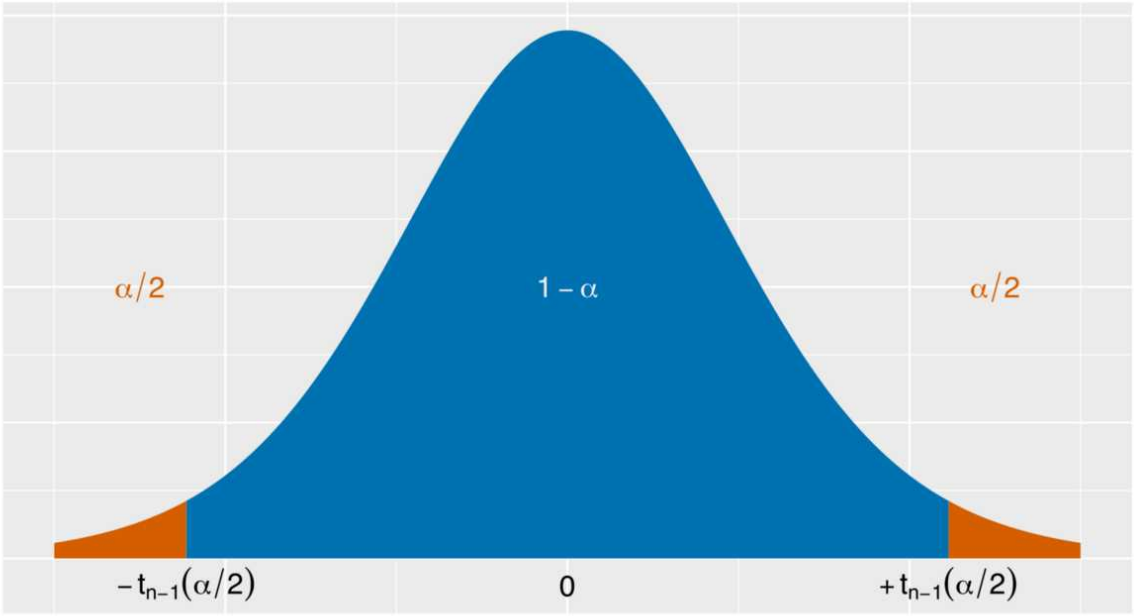
Sample means become more and more normally distributed around the true mean (the population parameter) as we increase our sample size. The variability of the sample means decreases as sample size increases.

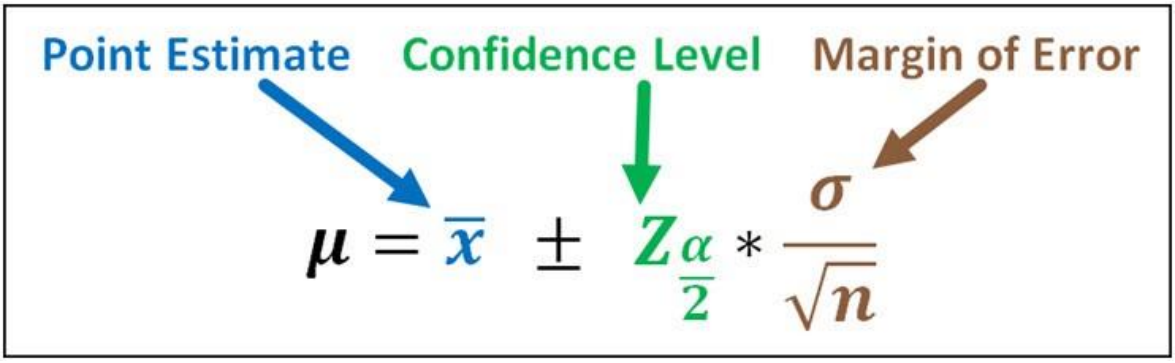
Central Limit Theorem

The Central Limit Theorem is used to help us understand the following facts regardless of whether the population distribution is normal or not:
1. the mean of the sample means is the same as the population mean
2. the standard deviation of the sample means is always equal to the standard error.
3. the *distribution* of *sample means* will become increasingly more *normal* as the sample size increases.

Confidence Intervals

A sample mean can be referred to as a point estimate of a population mean. Aconfidence interval is always centered around the mean of your sample. To construct the interval, you add a margin of error. The margin of error is found by multiplying the standard error of the mean by the z-score of the percent confidence level:

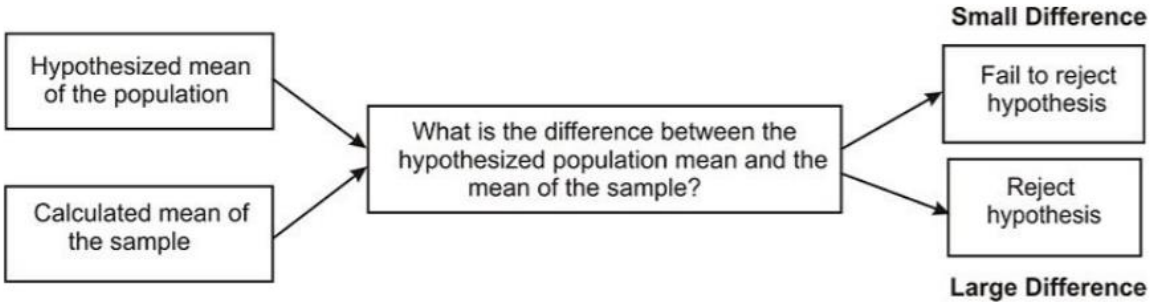




The confidence level indicates the number of times out of 100 that the mean of the population will be within the given interval of the sample mean.

Hypothesis Testing

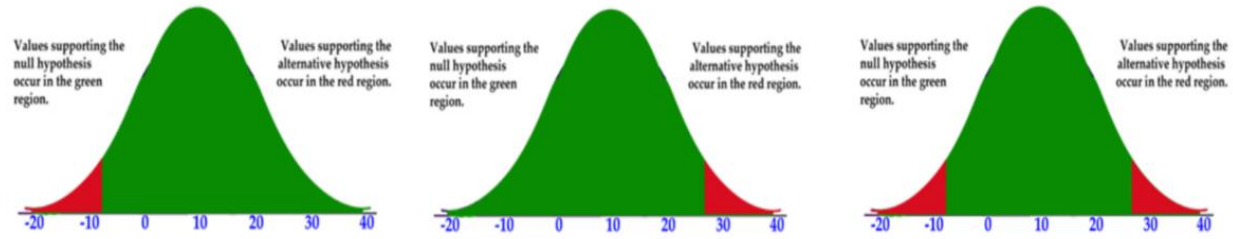
Hypothesis testing is a kind of statistical inference that involves asking a question, collecting data, and then examining what the data tells us about how to proceed. The hypothesis to be tested is called the null hypothesis and given the symbol *Ho*. We test the null hypothesis against an alternative hypothesis, which is given the symbol *Ha*.



When a hypothesis is tested, we must decide on how much of a difference between means is necessary in order to reject the null hypothesis. Statisticians first choose a level of significance or alpha(α) level for their hypothesis test.

Decision Made	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	Type I Error	Correct Decision
Do not Reject Null Hypothesis	Correct Decision	Type II Error

Critical values are the values that indicate the edge of the critical region. Critical regions describe the entire area of values that indicate you reject the null hypothesis.



left, right & two-tailed tests

These are the four basic steps we follow for (one & two group means) hypothesis testing:

1. State the null and alternative hypotheses.
2. Select the appropriate significance level and check the test assumptions.
3. Analyse the data and compute the test statistic.
4. Interpret the result.

Hypothesis Testing (One and Two Group Means)

Hypothesis Test on One Sample Mean When the Population Parameters are Known

We find the z-statistic of our sample mean in the sampling distribution and determine if that z-score falls within the critical(rejection) region or not. This test is only appropriate when you know the true mean and standard deviation of the population.

z-Test

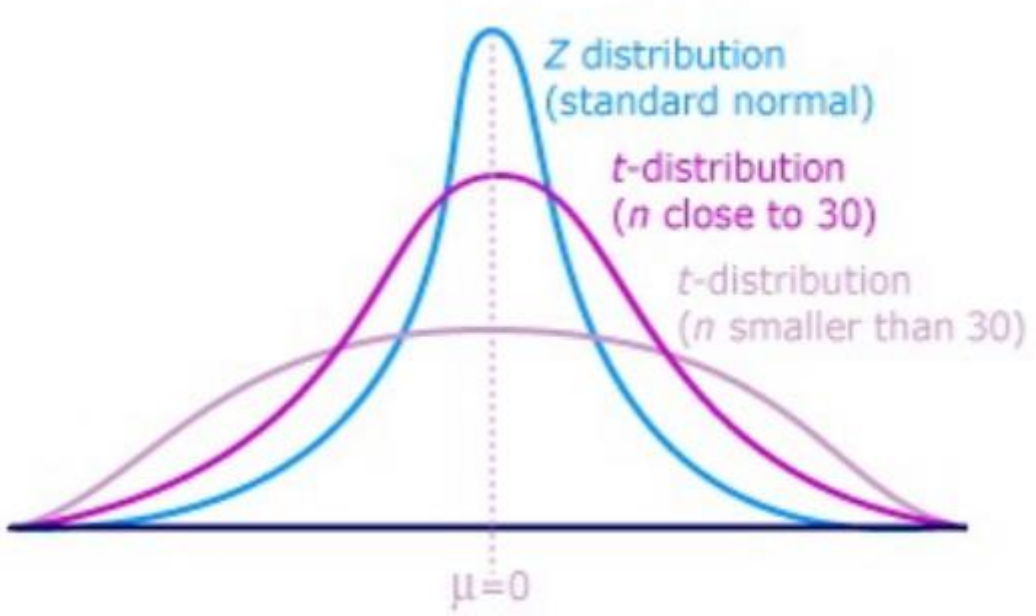
Formula to find the value of Z (z-test) Is:

$$Z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

$\bar{x}$ = mean of sample
μ_0 = mean of population
σ = standard deviation of population
n = no. of observations

Hypothesis Tests When You Don't Know Your Population Parameters

The Student's t-distribution is similar to the normal distribution, except it is more spread out and wider in appearance and has thicker tails. The differences between the t-distribution and the normal distribution are more exaggerated when there are fewer data points, and therefore fewer degrees of freedom.



t-Test

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

where:

t is the test statistic and has $n - 1$ degrees of freedom.

$\bar{x}$ is the sample mean

μ_0 is the population mean under the null hypothesis.

s is the sample standard deviation

n is the sample size

$\frac{s}{\sqrt{n}}$ is the estimated standard error

Estimation as a follow-up to a Hypothesis Test

When a hypothesis is rejected, it is often useful to turn to estimation to try to capture the true value of the population mean.

Two-Sample T Tests

Independent Vs Dependent Samples

When we have independent samples, we assume that the scores of one sample do not affect the other.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

$\bar{x}_1 - \bar{x}_2$ is the difference between the sample means

$\mu_1 - \mu_2$ is the difference between the hypothesized population means

$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is the standard error of the difference between the sample means

unpaired t-test

In two dependent samples of data, each score in one sample is paired with a specific score in the other sample.

$$t = \frac{\bar{d} - \delta}{SE_{\bar{d}}}$$

paired t-test

Hypothesis Testing (Categorical Data)

Chi-square test is used for categorical data and it can be used to estimate how closely the distribution of a categorical variable matches an expected distribution (the goodness-of-fit test), or to estimate whether two categorical variables are independent of one another (the test of independence).

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$

goodness-of-fit

degree of freedom (d f) = no. of categories(c)–1

$$expected\ cell\ value = \frac{C \times R}{n}$$

test of independence

degree of freedom (df) = (rows–1)(columns–1)

Hypothesis Testing (More Than Two Group Means)

Analysis of Variance (ANOVA) allows us to test the hypothesis that multiple population means and variances of scores are equal. We can conduct a series of t-tests instead of ANOVA but that would be tedious due to various factors.

We follow a series of steps to perform ANOVA:

1. Calculate the total sum of squares (SST)
2. Calculate the sum of squares between (SSB)
3. Find the sum of squares within groups (SSW) by subtracting
4. Next solve for degrees of freedom for the test
5. Using the values, you can now calculate the Mean Squares Between (MSB) and Mean Squares Within (MSW) using the relationships below
6. Finally, calculate the F statistic using the following ratio
7. It is easy to fill in the Table from here—and also to see that once the SS and df are filled in, the remaining values in the table for MS and F are simple calculations
8. Find F critical

$SS_T = \sum(y - \bar{y}^2) = \sum y^2 - \frac{(\sum y)^2}{N}$:

Where:
y = each observation
N = total number of scores
$\bar{y}$ = grand mean (mean of all scores)

$SS_B = \sum n_k (\bar{y}_k - \bar{y})^2$

Where:
k = the number of groups
n_k = the number of scores in group k
$\bar{y}_k$ = the mean of group k

$SS_W = SS_T - SS_B$

$df_{Total} = N - 1$
$df_{Between} = k - 1$
$df_{Within} = N - k$

$MS_B = \frac{SS_B}{df_{Between}}$

$MS_W = \frac{SS_W}{df_{Within}}$

$F = \frac{MS_B}{MS_W}$:

ANOVA formulas

If F-value from the ANOVA test is greater than the F-critical value, so we would reject our Null Hypothesis.

One-Way ANOVA

One-way ANOVA method is the procedure for testing the null hypothesis that the population means and variances of *a single independent variable* are equal.

Two-Way ANOVA

Two-way ANOVA method is the procedure for testing the null hypothesis that the population means and variances of two independent variables are equal. With this method, we are not only able to study

the effect of two independent variables, but also the interaction between these variables.

We can also do two separate one-way ANOVA but two-way ANOVA gives us Efficiency, Control & Interaction.

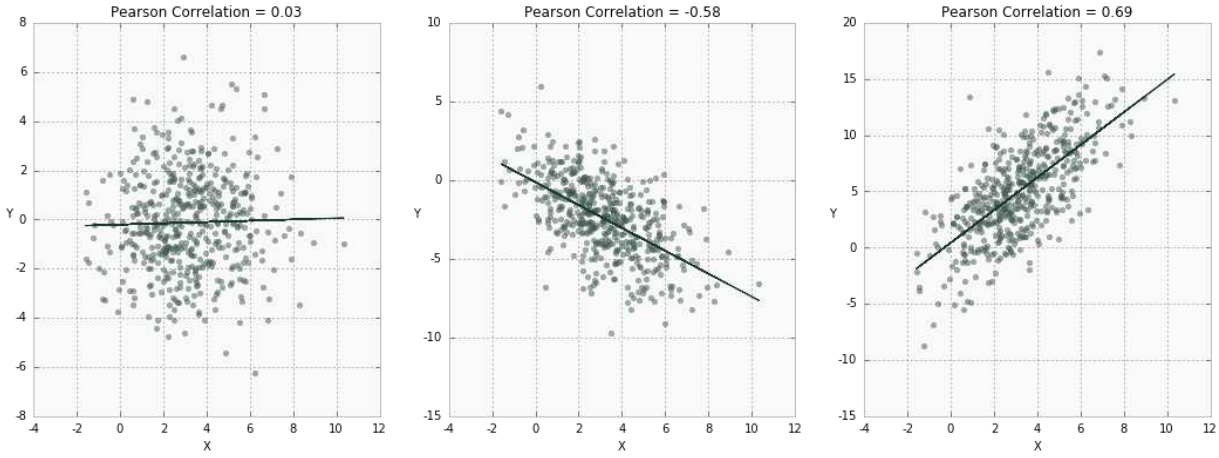
Quantitative Data (Correlation & Regression)

Correlation

Correlation refers to a mutual relationship or association between quantitative variables. It can help in predicting one quantity from another. It often indicates the presence of a causal relationship. It used as a basic quantity and foundation for many other modeling techniques.

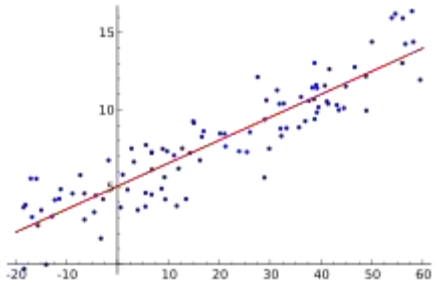
$$\rho_{X,\,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Pearson Correlation



Regression

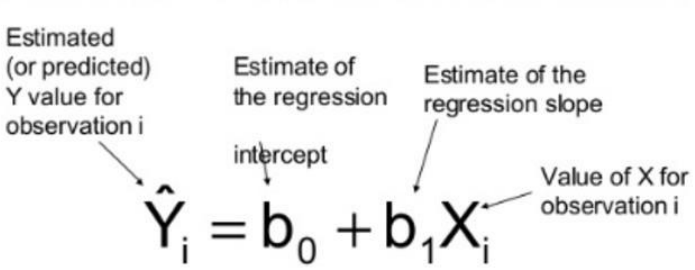
Regression analysis is a set of statistical processes for estimating the relationships among variables.

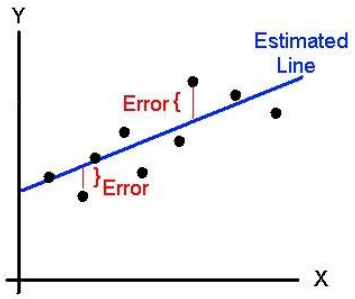


Regression

Simple Regression

This method uses a single independent variable to predict a dependent variable by fitting the best relationship.

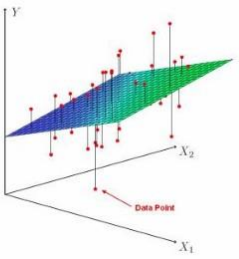




Multiple Regression

This method uses more than one independent variable to predict a dependent variable by fitting the best relationship.

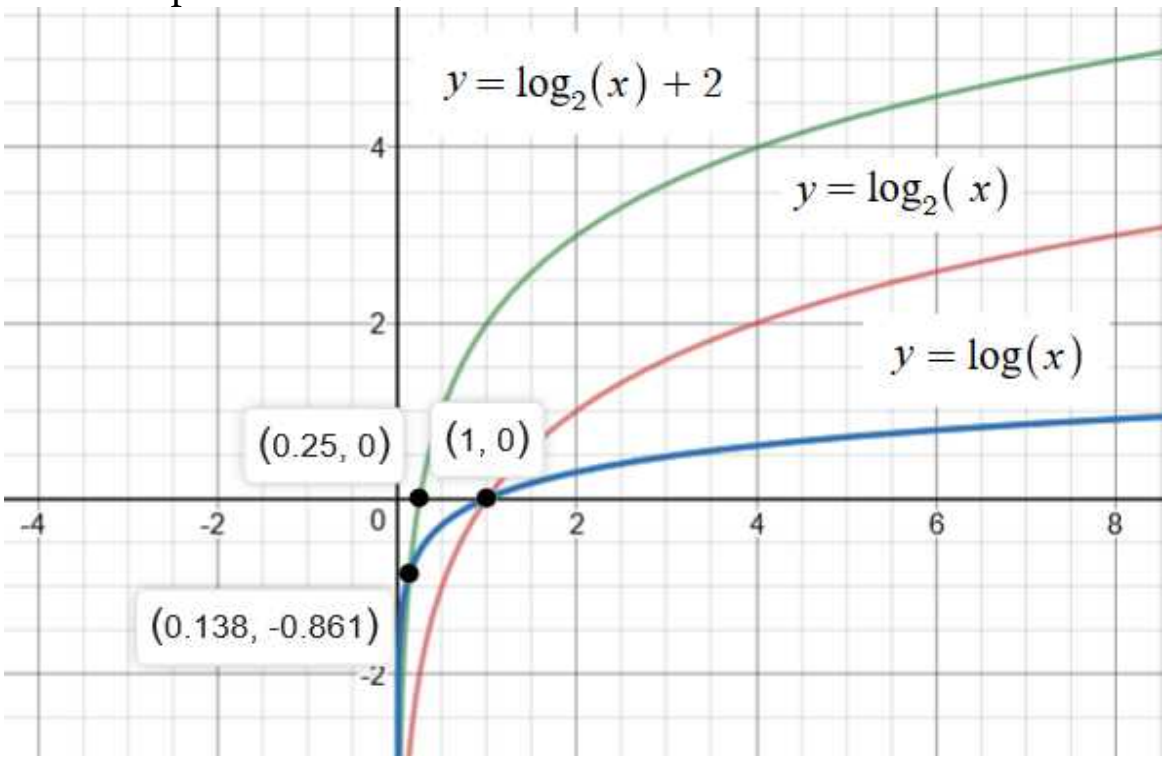
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n + \varepsilon$$



It works best when multicollinearity is absent. It's a phenomenon in which two or more predictor variables are highly correlated.

Nonlinear Regression

In this method, observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables.



Significance in Data Science

In data science, inferential statistics is used is many ways:

- Making inferences about the population from the sample.
- Concluding whether a sample is significantly different from the population.
- If adding or removing a feature from a model will really help to improve the model.
- If one model is significantly better than the other?
- Hypothesis testing in general.

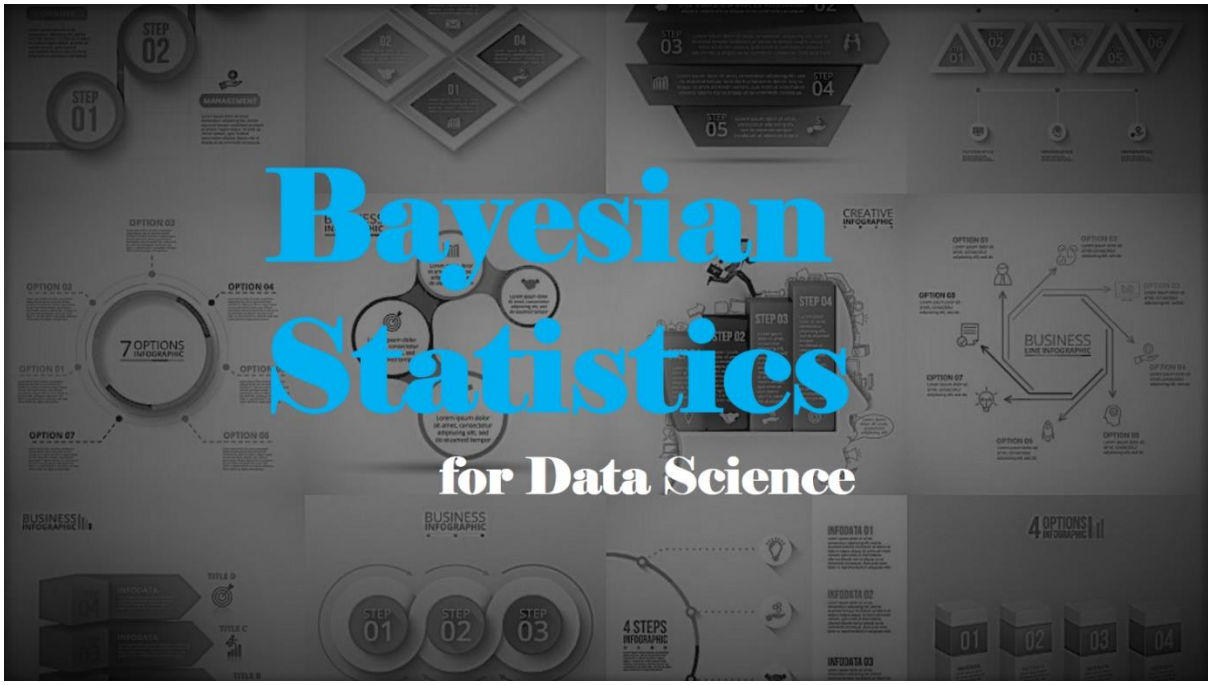
References:

Udacity
Inferential Statistics
classroom.udacity.com

Edx
Inferential Statistics
www.edx.org

Chapter 5

Bayesian Statistics



This chapter covers these topics related to Bayesian statistics and their significance in data science.

- *Frequentist Vs Bayesian Statistics*
- *Bayesian Inference*
- *Test for Significance*
- *Significance in Data Science*

Frequentist Vs Bayesian Statistics

Frequentist Statistics tests whether an event (hypothesis) occurs or not. It calculates the probability of an event in the long run of the experiment. A very common flaw found in frequentist approach i.e. dependence of the

result of an experiment on the number of times the experiment is repeated.

Frequentist statistics suffered some great flaws in its design and interpretationwhich posed a serious concern in all real life problems:

1. p-value & Confidence Interval (C.I) depend heavily on the sample size.
2. Confidence Intervals (C.I) are not probability distributions

Bayesian statistics is a mathematical procedure that applies probabilities to statistical problems. It provides people the tools to update their beliefs in the evidence of new data.

Frequentist Statistics	Bayesian Statistics
Parameters fixed	Parameters vary
Data varies	Data fixed
Probability P(D/Ø)	Likelihood P(Ø/D)
Confidence Interval	Credible Interval
No Prior	Strength of Prior

Frequentist Vs Bayesian Statistics

Frequentist vs. Bayesian Inference - The Basics of Bayesian Statistics | Coursera
This course describes Bayesian statistics, in which one's inferences about parameters or hypotheses are updated as...
www.coursera.org

Bayesian Inference

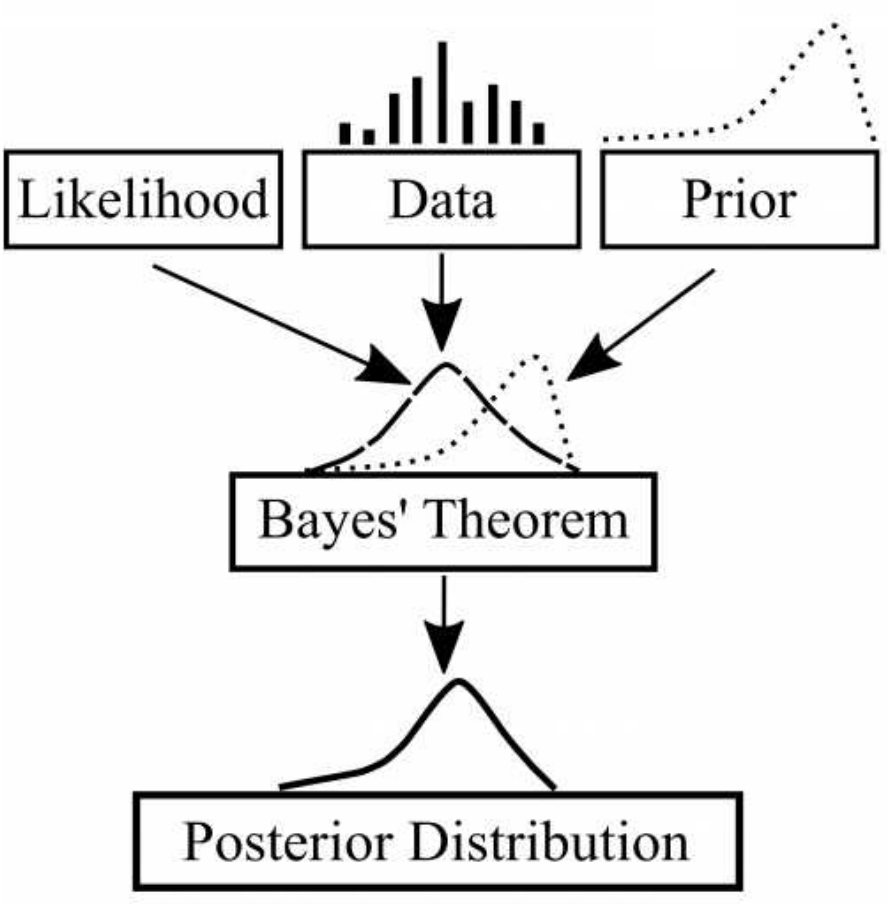
To understand Bayesian Inference, you need to understand Conditional Probability & Bayes Theorem, if you want to review these concepts, please refer my earlier post in this series.

Probability for Data Science
This is the 2nd post of blog post series 'Probability & Statistics for Data Science', this post covers these topics...
towardsdatascience.com

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

An important part of Bayesian Inference is the establishment of parameters andmodels. Models are the mathematical formulation of the observed events.Parameters are the factors in the models affecting the observed data. To define our model correctly , we need two mathematical models before hand. One to represent the likelihood function and the other for representing the distribution of prior beliefs **.** The product of these two gives the posterior belief distribution.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$



Courtesy: http://jason-doll.com/wordpress/?page_id=127

Likelihood Function

A likelihood function is a function of the parameters of a statistical model, given specific observed data. Probability describes the plausibility of a random outcome, without reference to any observed data while Likelihood describes the plausibility of a model parameter value, given specific observed data.

$$L(\emptyset/x) = cg(x/\emptyset)$$

L = Likelihood function
Ø = Parameters in probability model
x = Data
g = Probability density function
c = Constant

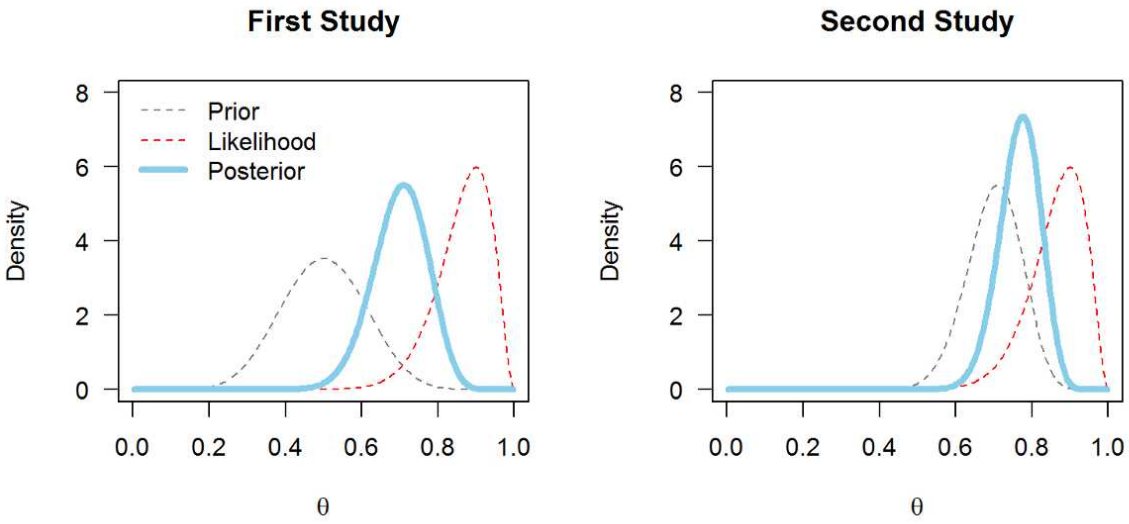
Likelihood function

Likelihood function - Wikipedia
In Bayesian inference, although one can speak about the likelihood of any proposition or random variable given another...
en.wikipedia.org

Prior & Posterior Belief distribution

Prior Belief distribution is used to represent our strengths on beliefs about the parameters based on the previous experience. Posterior Belief distribution is derived from multiplication of likelihood function & Prior Belief distribution.

As we collect more data, our posterior belief moves towards prior belief from likelihood:



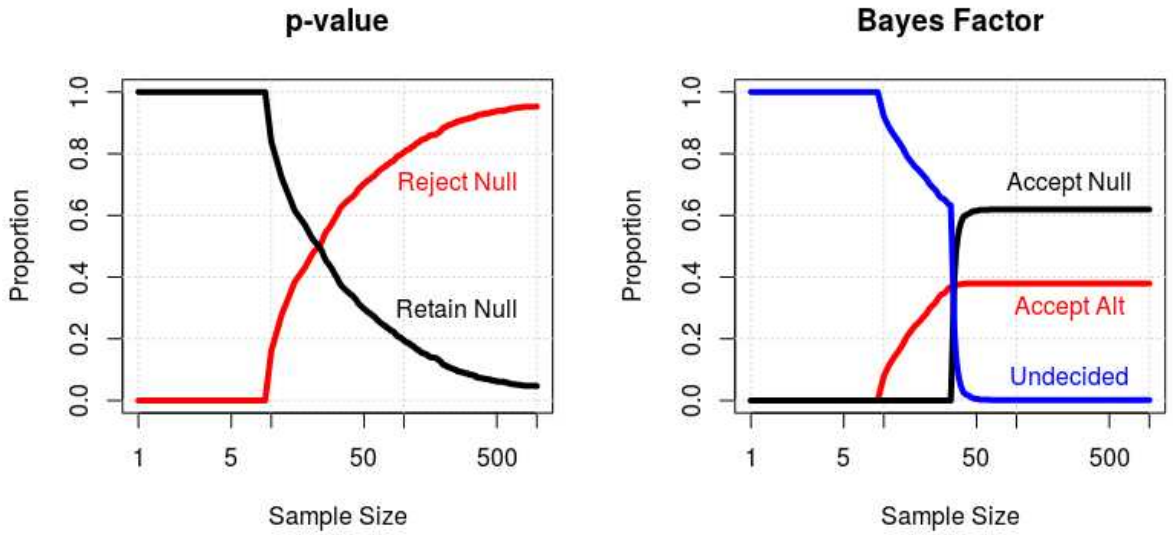
Courtesy: https://jimgrange.wordpress.com/2016/01/18/pesky-priors/

(Pesky?) Priors
When I tell people I am learning Bayesian statistics, I tend to get one of two responses: either people look at me...
jimgrange.wordpress.com

Test for Significance

Bayes factor

Bayes factor is the equivalent of p-value in the Bayesian framework. The null hypothesis in Bayesian framework assumes ∞ probability distribution only at a particular value of a parameter (say θ=0.5) and a zero-probability elsewhere. The alternative hypothesis is that all values of θ are possible, hence a flat curve representing the distribution.



Courtesy: http://areshenk-research-notes.com/bayes-factors-and-stopping-rules/

Using Bayes Factor instead of p-values is more beneficial in many cases since they are independent of intentions and sample size.

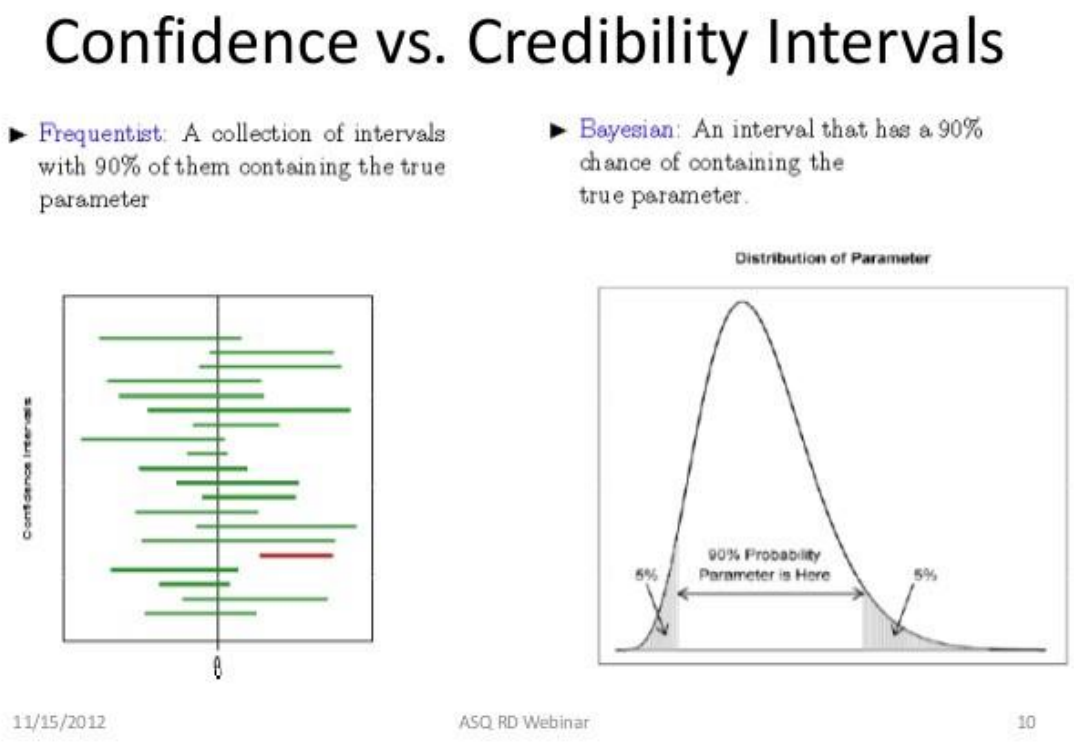
Replacing p-values with Bayes-Factors: A Miracle Cure for the Replicability Crisis in Psychological…

How Science Should Work Lay people, undergraduate students, and textbook authors have a simple model of science…

replicationindex.wordpress.com

High Density Interval (HDI)

High Density Interval (HDI) or Credibility Interval is equivalent to Confidence Interval (CI) in Bayesian framework. HDI is formed from the posterior distribution after observing the new data.



Courtesy: https://www.slideshare.net/ASQwebinars/bayesian-methods-in-reliability-engineering-15204318

Using High Density Interval (HDI) instead of Confidence Interval (CI) is more beneficial since they are independent of intentions and sample size.

Confidence vs. Credibility Intervals

Tomorrow, for the final lecture of the Mathematical Statistics course, I will try to illustrate - using Monte Carlo...

freakonometrics.hypotheses.org

Moreover, there is a nice article published on *AnalyticsVidhya* on this which elaborate on these concepts with examples:

Bayesian Statistics explained to Beginners in Simple English

Introduction Bayesian Statistics continues to remain incomprehensible in the ignited minds of many analysts. Being...

www.analyticsvidhya.com

Significance in Data Science

Bayesian statistics encompasses a specific class of models that could be used for Data Science. Typically, one draws on Bayesian models for one or more of a variety of reasons, such as:

- having relatively few data points
- having strong prior intuitions
- having high levels of uncertainty

And there are scenarios where Bayesian statistics will perform drastically, please read following discussion for details:

What's the relationship between bayesian statistics and machine learning?
Answer (1 of 2): Machine learning is a broad field that uses statistical models and algorithms to automatically learn...
www.quora.com

References:

Bayesian Methods for Hackers
The Bayesian method is the natural approach to inference, yet it is hidden from readers behind chapters of slow...
nbviewer.jupyter.org

Bayesian Statistics: From Concept to Data Analysis | Coursera
About this course: This course introduces the Bayesian approach to statistics, starting with the concept of probability...
www.coursera.org

Chapter 6

Statistical Learning

This chapter covers these topics related to Statistical Learning and their significance in data science.

- *Introduction*
- *Prediction & Inference*
- *Parametric & Non-parametric methods*
- *Prediction Accuracy and Model Interpretability*
- *Bias-Variance Trade-Off*

Introduction

Statistical learning is a framework for understanding data based on statistics, which can be classified as supervised or unsupervised. Supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs, while

in unsupervised statistical learning, there are inputs but no supervising output; but we can learn relationships and structure from such data.

One of the simple way to understand statistical learning is to determine association between predictors (independent variables, features) & response(dependent variable) and developing a accurate model that can predict response variable (*Y*) on basis of predictor variables (*X*).

Y = f(X) + ε where X = (X1,X2, . . .,Xp), f is an *unknown function* & ε is *random error (reducible & irreducible).*

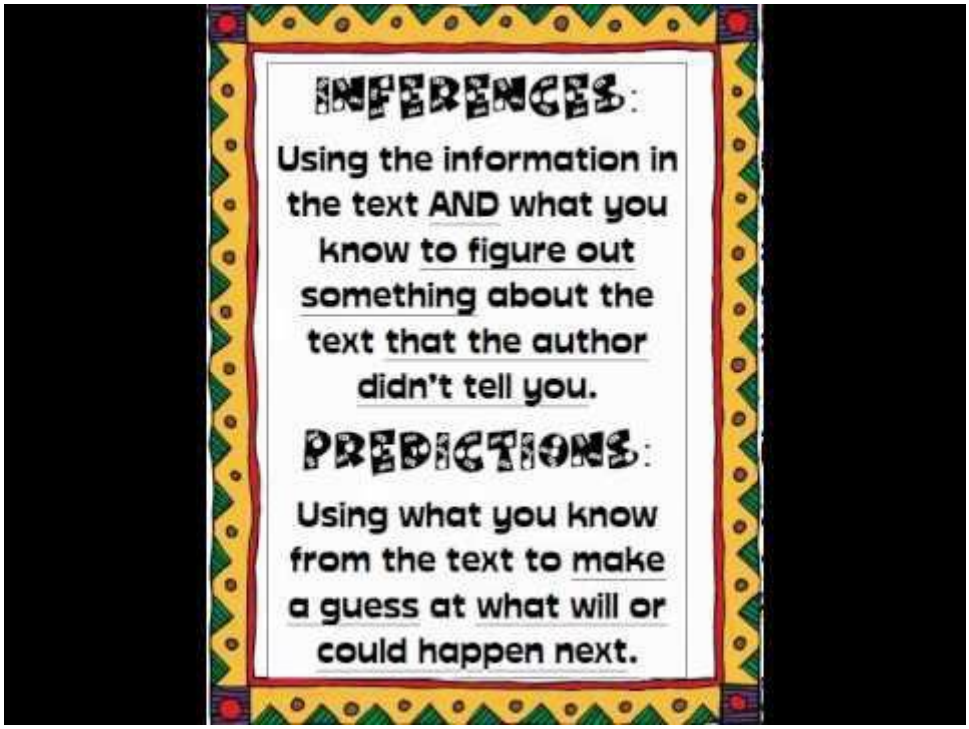
Introduction to Statistical Learning
"As a former data scientist, there is no question I get asked more than, "What is the best way to learn statistics?" I...
www-bcf.usc.edu

Prediction & Inference

In situations where a set of inputs X are readily available, but the output Y is not known, we often treat f as black box (not concerned with the exact form of f), as long as it yields accurate predictions for Y . This is prediction.

There are situations where we are interested in understanding the way that Y is affected as X change. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y . Here we are more interested in understanding relationship between X and Y. Now f cannot be treated as a black box, because we need to know its exact form. This is inference.

In real life, will see a number of problems that fall into the prediction setting, the inference setting, or a combination of the two.



Courtesy: https://www.youtube.com/watch?v=w09Ifi62p8k

What is the difference between an inference and a prediction?
Answer (1 of 5): In both the cases, you have some input variables and then you have the response on the other side...
www.quora.com

Parametric & Non-parametric methods

When we assume about the functional form of f and try to estimate f by estimating the set of parameters, these methods are called parametric methods.

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Non-parametric methods do not make explicit assumptions about the form of f, instead they seek an estimate of f that gets as close to the data points as possible.

Parametric vs Nonparametric Models

Terminology (roughly):

- **Parametric Models** have a finite fixed number of parameters $\boldsymbol{\theta}$, regardless of the size of the data set. Given $\boldsymbol{\theta}$, the predictions are independent of the data $\mathcal{D}$:

$$p(x, \boldsymbol{\theta}|\mathcal{D}) = p(x|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathcal{D})$$

  The parameters are a finite summary of the data. We can also call this model-based learning (e.g. mixture of k Gaussians)

- **Non-parametric Models** allow the number of "parameters" to grow with the data set size, or alternatively we can think of the predictions as depending on the data, and possible a usually small number of parameters $\boldsymbol{\alpha}$

$$p(x|\mathcal{D}, \alpha)$$

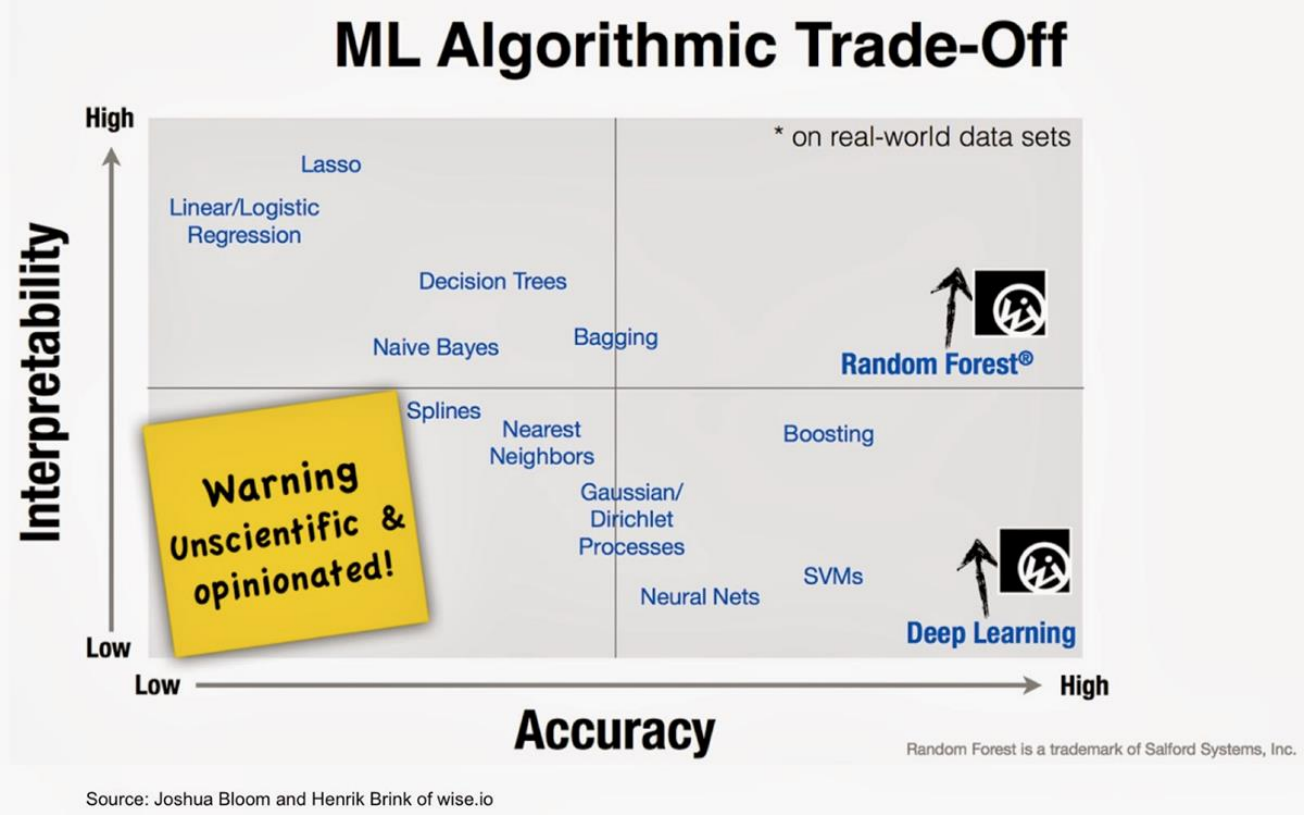
  We can also call this memory-based learning (e.g. kernel density estimation)

Courtesy: https://www.slideshare.net/zukun/icml2004-tutorial-on-bayesian-methods-for-machine-learning

Parametric and Nonparametric Machine Learning Algorithms
What is a parametric machine learning algorithm and how is it different from a nonparametric machine learning...
machinelearningmastery.com

Prediction Accuracy and Model Interpretability

Of the many methods that we use for statistical learning, some are less flexible, or more restrictive. When inference is the goal, there are clear advantages to using simple and relatively inflexible statistical learning methods. When we are only interested in prediction, we use flexible models available.



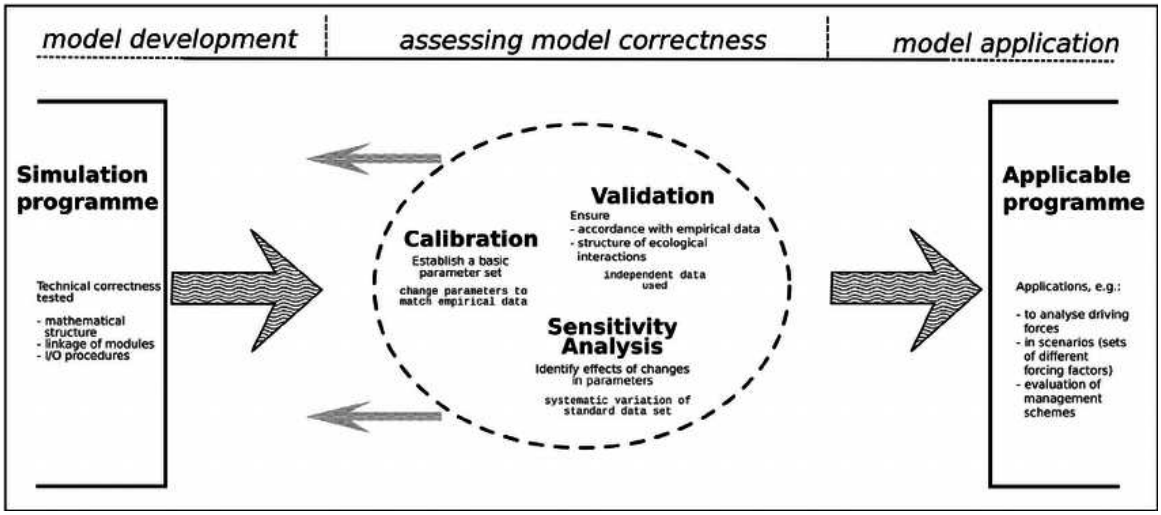
Courtesy: https://towardsdatascience.com/a-complete-machine-learning-walk-through-in-python-part-three-388834e8804b

Model Prediction Accuracy Versus Interpretation in Machine Learning
In their book , Kuhn and Johnson comment early on the trade-off of model prediction accuracy versus model...
machinelearningmastery.com

Predictive modeling: Striking a balance between accuracy and interpretability
Editor's note: Register for the free webcast "How the machine learning wave is changing the way organizations look at...
www.oreilly.com

Assessing Model Accuracy

There is no free lunch in statistics, which means no one method dominates all others over all possible data sets. In the regression setting, the most commonly-used measure is the mean squared error (MSE). In the classification setting, the most commonly-used measure is the confusion matrix. Fundamental property of statistical learning is that, as model flexibility increases, training error will decrease, but the test error may not.



Courtesy: https://www.researchgate.net/figure/Basic-steps-to-assess-model-accuracy-Assuming-a-technically-approved-model-the-next-step_fig1_276365016

How to Evaluate Machine Learning Algorithms

Once you have defined your problem and prepared your data you need to apply machine learning algorithms to the data in…

machinelearningmastery.com

7 Important Model Evaluation Error Metrics Everyone should know

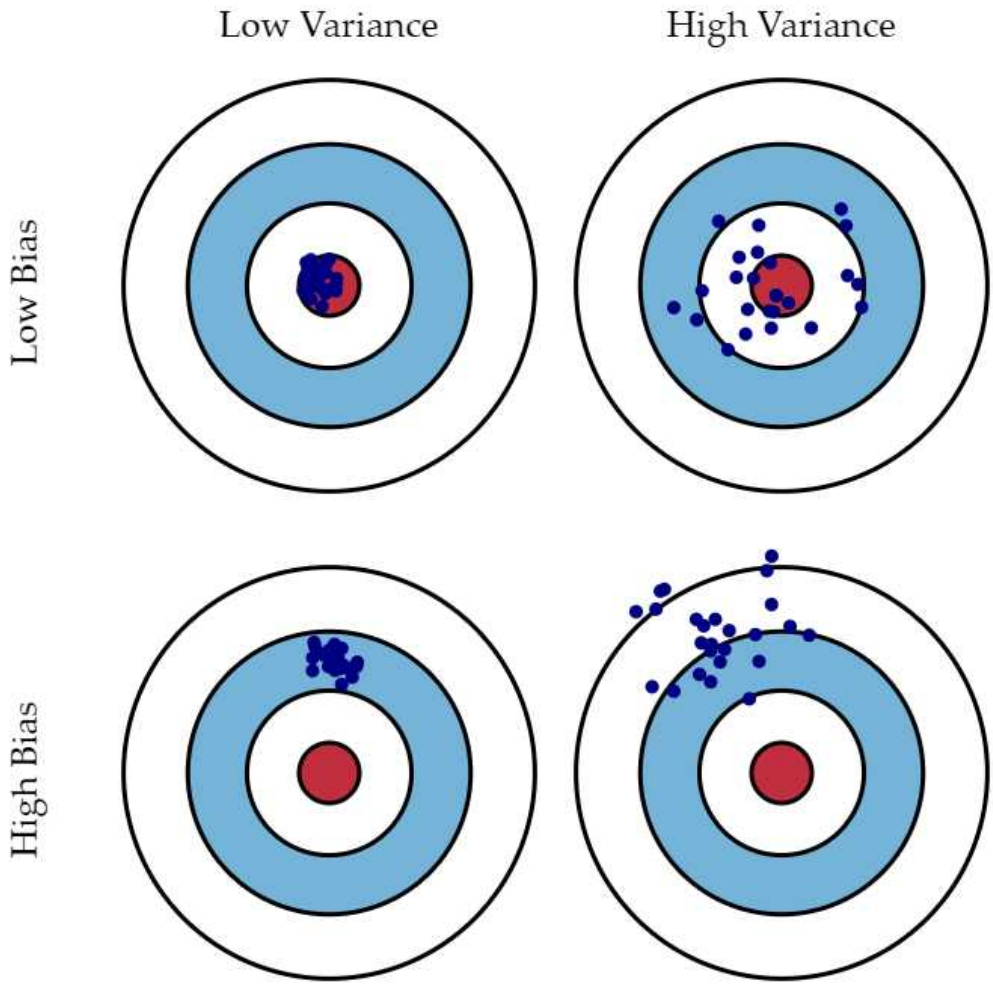
Introduction Predictive Modeling works on constructive feedback principle. You build a model. Get feedback from…

www.analyticsvidhya.com

Bias & Variance

Bias are the simplifying assumptions made by a model to make the target function easier to learn. Parametric models have a high bias making them fast to learn and easier to understand but generally less flexible. Decision Trees, k-Nearest Neighbours and Support Vector Machines are low-bias machine learning algorithms. Linear Regression, Linear Discriminant Analysis and Logistic Regression are high-bias machine learning algorithms.

Variance is the amount that the estimate of the target function will change if different training data was used. Non-parametric models that have a lot of flexibility have a high variance. Linear Regression, Linear Discriminant Analysis and Logistic Regression are low-variance machine learning algorithms. Decision Trees, k-Nearest Neighbours and Support Vector Machines are high-variance machine learning algorithms.



Courtesy: http://scott.fortmann-roe.com/docs/BiasVariance.html

What is the best way to explain the bias-variance trade-off in layman's terms ?
Answer (1 of 7): This picture should be a great way to explain bias-variance to a 5-year-old. For the group of smart...
www.quora.com

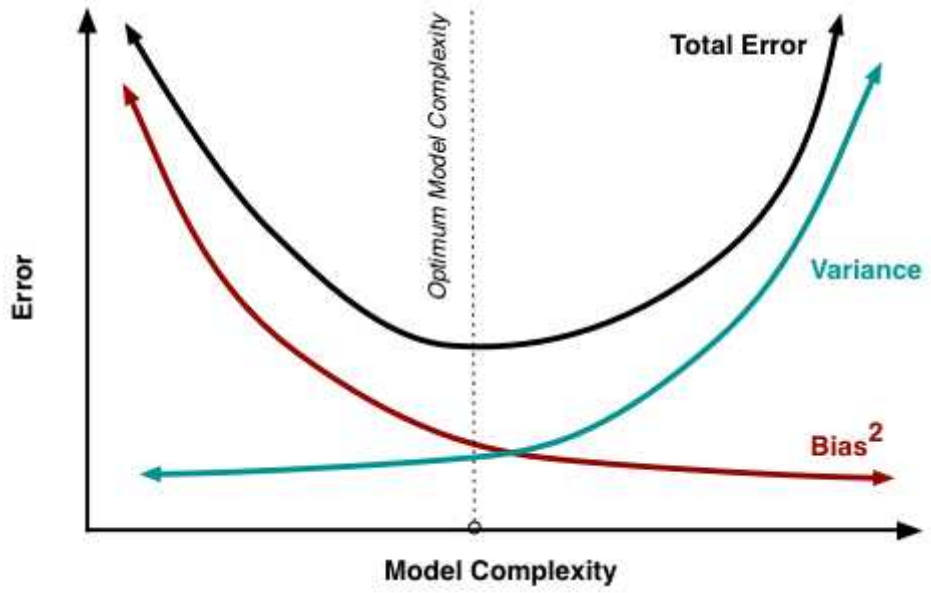
Bias-Variance Trade-Off

The relationship between bias and variance in statistical learning is such that:

- Increasing bias will decrease variance.
- Increasing variance will decrease bias.

There is a trade-off at play between these two concerns and the models we choose and the way we choose to configure them are finding different balances in this trade-off for our problem.

In both the regression and classification settings, choosing the correct level of flexibility is critical to the success of any statistical learning method. The bias-variance trade-off, and the resulting U-shape in the test error, can make this a difficult task.



Courtesy: http://scott.fortmann-roe.com/docs/BiasVariance.html

Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning
Supervised machine learning algorithms can best be understood through the lens of the bias-variance trade-off. In this...
machinelearningmastery.com

Understanding the Bias-Variance Tradeoff
Whenever we discuss model prediction, it's important to understand prediction errors (bias and variance). There is a...
towardsdatascience.com