



STATISTICAL FOUNDATIONS



Over View

- ▶ Descriptive Statistics
 - ▶ Data Representation
 - ▶ Graphical Representation
 - ▶ Tabular Representation
 - ▶ Summary Statistics
 - ▶ Probability Distribution & Random Variables
- ▶ •Inferential Statistics
 - ▶ Confidence Intervals
 - ▶ Hypothesis Testing
 - ▶ Estimating Parameters to establish Relations



Statistics - Introduction

▶ **What is Statistics?**

- ▶ *It is defined as the science, which deals with the collection, analysis and interpretation of data.*

▶ **Scope of Statistics:** It find applications in almost all possible areas like Planning, Economics, Business, Biology, Astronomy, Medical Science, Psychology, Education and even in War.

▶ **Limitations of Statistics:** The following are some of its important limitations:

- ▶ *Statistics does not study individuals.*
- ▶ *Statistical laws are not exact.*
- ▶ *Statistics is liable to be misused*



Use of Data Analytics

- ▶ **Man vs Machine**
 - ▶ IBM Deep Blue Beats Grand Master Gary Kasparov (1997).
 - ▶ Deep Blue train itself using historical Chess Game Data to train the software
- ▶ **Machine vs Machine**
 - ▶ Deep Mind's Alphazero beats the best chess engine of the time Stock fish(2017)
 - ▶ Alphazero uses Reinforcement learning to train itself by just using the set of rules of chess.
- ▶ **Analytics in Games**
 - ▶ Oakland Athletics Baseball team Manager Billy Beane develop **Sabermetrics**
 - ▶ Saber metricians collect in game activity data to take key strategic decisions during the game



Descriptive Statistics

- ▶ describing, presenting, summarizing, and organizing your data, either through numerical calculations or graphs or tables.
- ▶ Some of the common measurements in descriptive statistics are central tendency and others the variability of the dataset.
- ▶ helps us to understand our data and is very important part of Machine Learning.
- ▶ Doing a descriptive statistical analysis of our dataset is absolutely crucial.



Descriptive Statistics

- ▶ • **What is Data?**
 - ▶ **Data are individual pieces of factual information recorded and used for the purpose of analysis.**
- ▶ Data is broadly classified into
 - ▶ **Quantitative Data: Numerical values**
 - ▶ Continuous
 - ▶ Discrete
 - ▶ **Qualitative Data: Categorical-Data is a group into discrete groups**
 - ▶ Nominal: Order does not exist
 - Ex: marital Status: Married/Unmarried
 - ▶ Ordinal: Order does exist
 - Ex: Player contract: A Class, B Class, C Class, D Class
- ▶ Vary High
- ▶ Skill player
- ▶ Significantly Low
- ▶ Skill player



Data Representation

- ▶ Data can be represented in following ways
 - ▶ Graphical Representation
 - ▶ Categorical Variable
 - ▶ Bar Chart
 - ▶ Pie Chart
 - ▶ Quantitative Variable
 - ▶ Box & Whisker Plot
 - ▶ Histogram Plot
 - ▶ Scatter Plot
 - ▶ Tabular Representation
 - ▶ Contingency Table



Summary Statistics

▶ Measure of Central Tendency:

- ▶ Mean
- ▶ Median
- ▶ Mode
- ▶ Quartile

▶ Measure of Statistical Dispersion:

- ▶ Mean Absolute Deviation (MAD)
- ▶ Standard Deviation (SD)
- ▶ Variance
- ▶ Inter Quartile Range (IQR)
- ▶ Range

▶ Measure of the Shape of the Distribution

- ▶ skewness
- ▶ kurtosis

▶ Measure of Statistical Dependence

- ▶ Covariance
 - ▶ Correlation Coefficient
-



Measure of Central Tendency

- ▶ It describes a whole set of data with a single value that represents the centre of its distribution. There are three main measures of central tendency:



Summary Statistics: Measure of Central Tendency - Mean

- ▶ Mean: For a data set, the Mean is a central value of a finite set of numbers.
- ▶ The arithmetic mean of a set of numbers x_1, x_2, \dots, x_n is typically denoted by \bar{x}

$$E[x] = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ▶ If the data set were based on a series of observations obtained by sampling from a statistical population, the arithmetic mean is the **sample mean**.



Summary Statistics: Measure of Central Tendency - Median

- ▶ Median: is the value separating the higher half from the lower half of a data.
- ▶ For a data set, it may be thought of as "the middle" value.
- ▶ In order to find Median we need to first sort the data points in increasing order and then Pick the middle element as Median.
- ▶ If number of data points are even then Median is average of middle pair

$$\text{Median}(X) = \begin{cases} x_{[n/2]} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{n/2+1}}{2} & \text{if } n \text{ is even} \end{cases}$$



-
- ▶ Example: $X=[3,4,3,1,2,3,9,5,6,7,4,8]$
 - ▶ Step 1: sort the data points we get $[1,2,3,3,3,4,4,5,6,7,8,9]$, here number of data points are even, $n=12$.
 - ▶ Step 2: Since n is even, median is average of middle pair.
 - ▶ Median(X) is 4.



Summary Statistics: Measure of Central Tendency - Mode

- ▶ Mode: is the value that appears most frequently in the dataset. It is the value in the dataset whose frequency is maximum.
 - ▶ If the dataset is of discrete values finding the Mode is the easier task.
 - ▶ If the dataset have continuous or real values finding Mode is not a easy task.
 - ▶ In case of real value dataset Mode is obtained with the help of Histogram.
 - ▶ In practice the mid point of the bin with highest frequency is consider as Mode.
 - ▶ The issue with this approach is as you change the bin size (or number of bins) the Mode will change.
 - ▶ This is the reason mode is not very popular in data analytics.
-



Summary Statistics: Measure of Central Tendency - Quartile

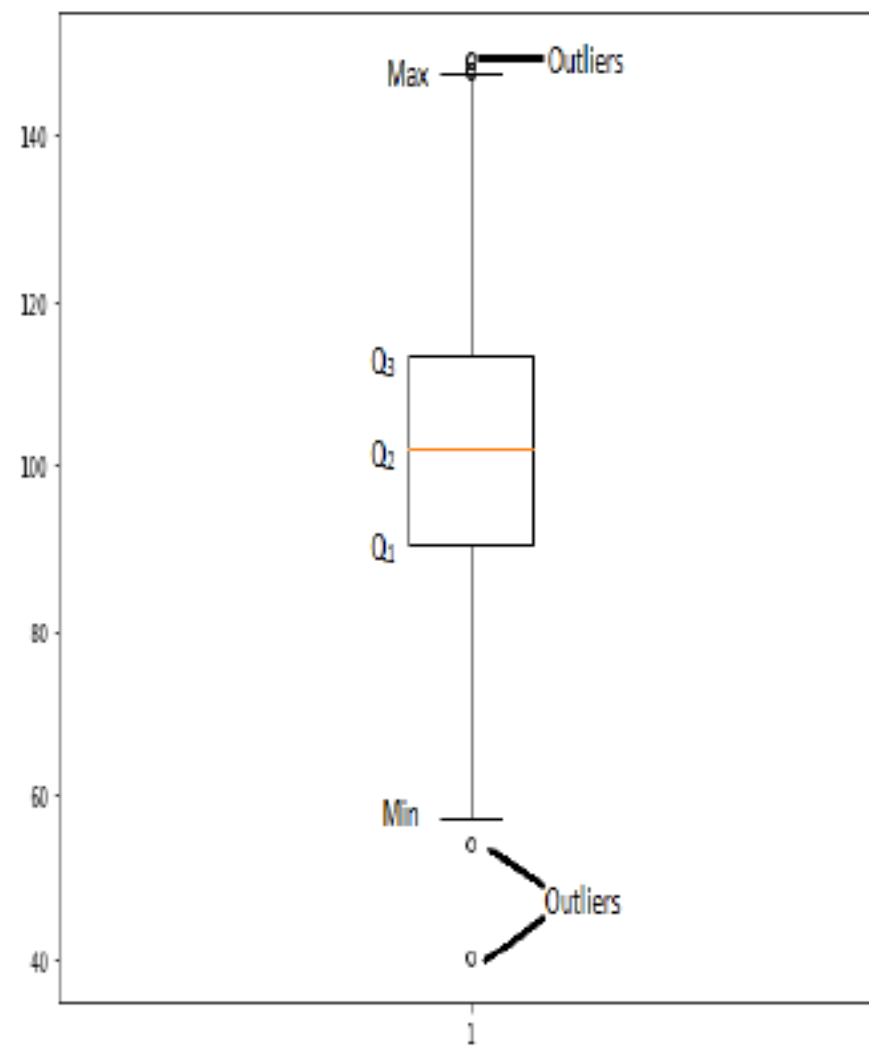
- ▶ Quartiles: are the points(values) which divides the data in quarters.
 - ▶ Quartile Q1 is the points(values) below which there are 25% data points of the dataset.
 - ▶ Quartile Q2 is the points(values) below which there are 50% data points of the dataset. Which is also the Median.
 - ▶ Quartile Q3 is the points(values) below which there are 75% data points of the dataset.
 - ▶ Sometimes the minimum value (after excluding outliers on lower side) is consider as Quartile Q0 and maximum value (after excluding outliers on higher side) of the Quartile Q5.
 - ▶ Outliers are all the datapoint which outside $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ range, where IQR is Inter Quartile Range.
-



Box & Whisker Plot

- ▶ Method for graphically depicting groups of numerical data through their quartiles.
- ▶ Box&Whisker plot display variation in samples of the population without making any assumptions of the underlying distribution
- ▶ Method of presenting the dataset based on a five point summary:
 - ▶ **Minimum:** the smallest value without outliers (if present in the dataset).
 - ▶ **Maximum:** the largest value without outliers (if present in the dataset).
 - ▶ **Median(Q2):** the middle value of the dataset.
 - ▶ **First quartile(Q1):** It is the median of the first half of the dataset.
 - ▶ **Third quartile(Q3):** It is the median of the second half of the dataset.





Box & Whisker Plot

Summary Statistics: Measure of Statistical Dispersion

- ▶ Inter Quartile Range (IQR): is the range in which the middle 50% data points of the data set belongs

$$IQR = Q_3 - Q_1$$

- ▶ Mean Absolute deviation (MAD) proposed by Gauss (1821)

$$MAD(x) = \frac{1}{n-1} \sum_{i=1}^n |x_i - \bar{x}|$$

- ▶ Standard deviation: the term was introduced by Karl Pearson in 1893

$$SD(x) = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$



Summary Statistics: Measure of Statistical Dispersion

- ▶ Mean Square Deviation : also known as Variance, proposed by Sir Ronald Aylmer Fisher in 1920

$$\begin{aligned} Var(x) &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= E[x^2] - (E[x])^2 \end{aligned}$$

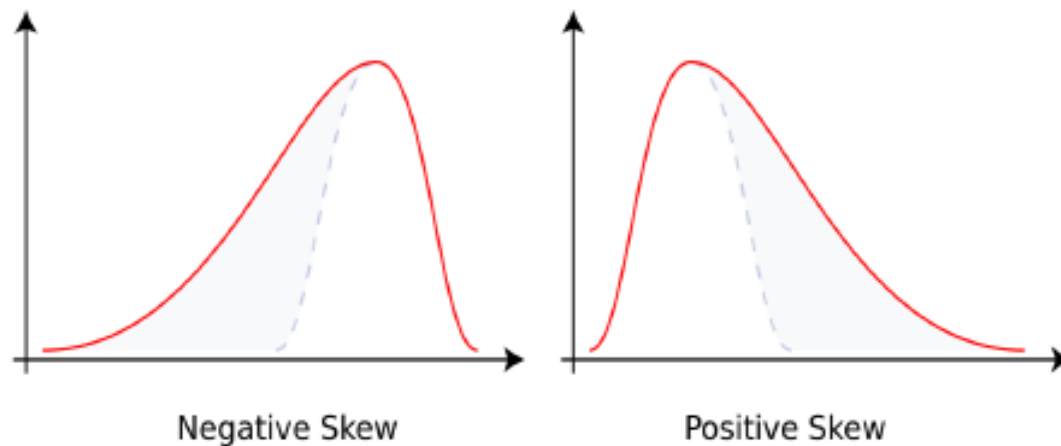
- ▶ Range : is the difference between the largest and smallest values.

$$Range(x) = \max(x) - \min(x)$$



Summary Statistics: Measure of the Shape of the Distribution - Skewness

- ▶ Skewness: is a measure of the asymmetry of the distribution of a random variable about its mean.
- ▶ •The skewness value can be positive, zero, negative.
- ▶ •Skewness of the distribution with tail on the right side is positive, and for left tail distribution it is negative. Skewness of the symmetric distribution is zero.



$$\text{Skewness}(x) = E \left[\left(\frac{x - \mu_x}{\sigma_x} \right)^3 \right] = \frac{\mu_3}{\sigma^3}$$

- For a sample of n values, estimators of the population skewness is

$$\text{Skewness}(x) = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$



Summary Statistics: Measure of the Shape of the Distribution -

- ▶ Kurtosis: is a measure of the "tailedness" of the probability distribution.
- ▶ •It is a scaled version of the fourth moment of the distribution.
- ▶ •Kurtosis of Normal distribution is Three.
- ▶ •Distribution with Kurtosis less than three are referred as **Platykurtic**, such distributions have less extreme outliers than Normal distribution. Example: Uniform distribution.
- ▶ •Distribution with Kurtosis greater than three are referred as **leptokurtic**, such distributions have more extreme outliers than Normal distribution .Example :Laplace distribution.



$$Kurtosis(x) = E\left[\left(\frac{x - \mu_x}{\sigma_x}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right]^2}$$

► $Kurtosis(x) \geq (Skewness(x))^2 + 1$



Summary Statistics: Measure of Statistical Dependence

- ▶ Covariance: is a measure of the joint variability of two random variables.
- ▶ If higher value of first variable corresponds to higher value of second variable and the lower value of first variable corresponds to lower value of second variable then the covariance between the pair of variables is **positive**. Reverse of that have covariance value negative.
- ▶ If the variables are independent then Covariance value is zero.

$$COV(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

- ▶ •Sample Covariance

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

Summary Statistics: Measure of Statistical Dependence

- ▶ Properties of covariance:
- ▶ Bilinear property: for a and b as a constant and random variable X,Y and Z,

$$COV((aX + bY), Z) = aCOV(X, Z) + bCOV(Y, Z)$$

Symmetric Property: $COV(X, Y) = COV(Y, X)$

Positive semi definite Property: $COV(X, X) = \sigma_X^2$

$$|COV(X, Y)| \leq \sigma_X \sigma_Y$$



Summary Statistics: Measure of Statistical Dependence

- ▶ Correlation coefficient: between pair of random variable X and Y with expected value μ_X and μ_Y and standard deviation σ_X and σ_Y is defined as

$$\rho_{X,Y} = \text{corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

- ▶ Value of Correlation coefficient is bounded between -1 and 1

$$-1 \leq \text{corr}(X,Y) \leq 1$$



Outlier Analysis

- ▶ Outlier – data objects that are grossly different from or inconsistent with the remaining set of data
- ▶ Causes
 - ▶ Measurement / Execution errors
 - ▶ Inherent data variability
- ▶ Outliers – maybe valuable patterns
 - ▶ Fraud detection
 - ▶ Customized marketing
 - ▶ Medical Analysis



Outlier Mining

- ▶ Given n data points and k – expected number of outliers find the top k dissimilar objects
- ▶ Define inconsistent data
 - ▶ Residuals in Regression
 - ▶ Difficulties – Multi-dimensional data, non-numeric data
- ▶ Mine the outliers
 - ▶ Visualization based methods
 - ▶ Not applicable to cyclic plots, high dimensional data and categorical data



Approaches

- ▶ Statistical Approach
- ▶ Distance-based approach
- ▶ Density based outlier approach
- ▶ Deviation based approach



► Variants of Outlier Detection Problems

- Given a database D , find all the data points $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
- Given a database D , find all the data points $\mathbf{x} \in D$ having the top- n largest anomaly scores $f(\mathbf{x})$
- Given a database D , containing mostly normal (but unlabeled) data points, and a test point \mathbf{x} , compute the anomaly score of \mathbf{x} with respect to D



Applications

- ▶ Credit card fraud detection
- ▶ telecommunication fraud detection
- ▶ network intrusion detection
- ▶ fault detection



▶ Challenges

- ▶ How many outliers are there in the data?
- ▶ Method is unsupervised
 - ▶ Validation can be quite challenging (just like for clustering)
- ▶ Finding needle in a haystack

▶ Working assumption:

- ▶ There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data



▶ General Steps

▶ Build a profile of the “normal” behavior

- ▶ Profile can be patterns or summary statistics for the overall population

▶ Use the “normal” profile to detect anomalies

- ▶ Anomalies are observations whose characteristics differ significantly from the normal profile

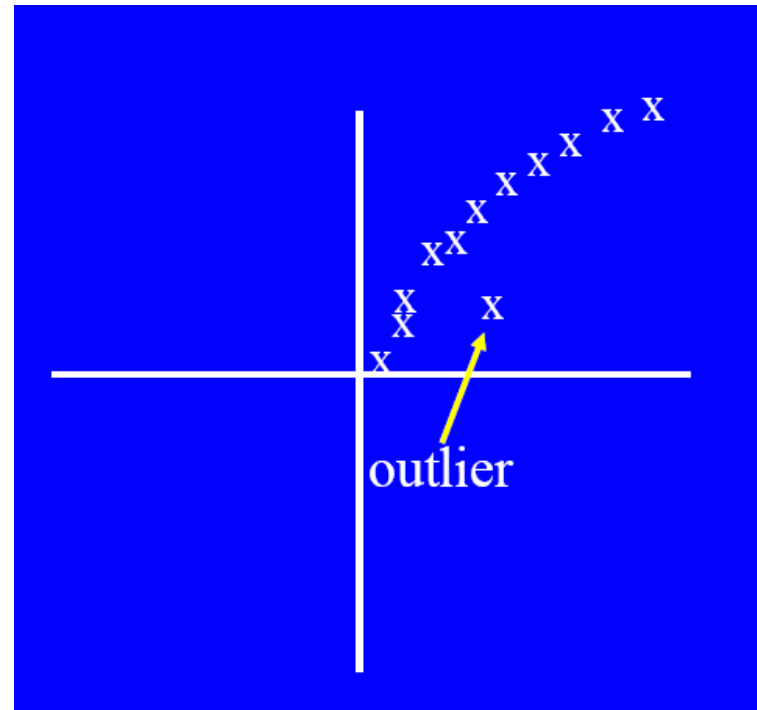
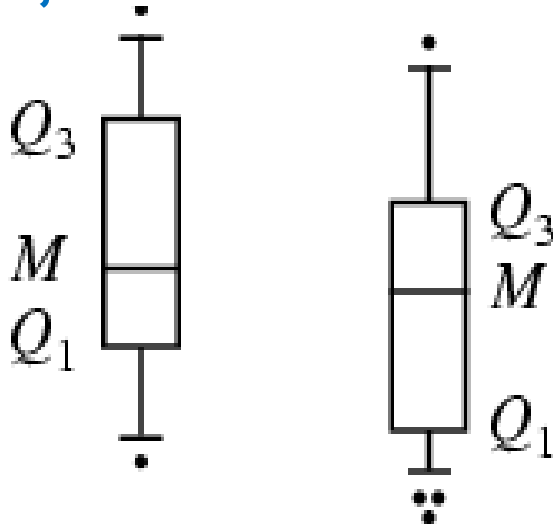


Graphical Approaches

- ▶ Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)

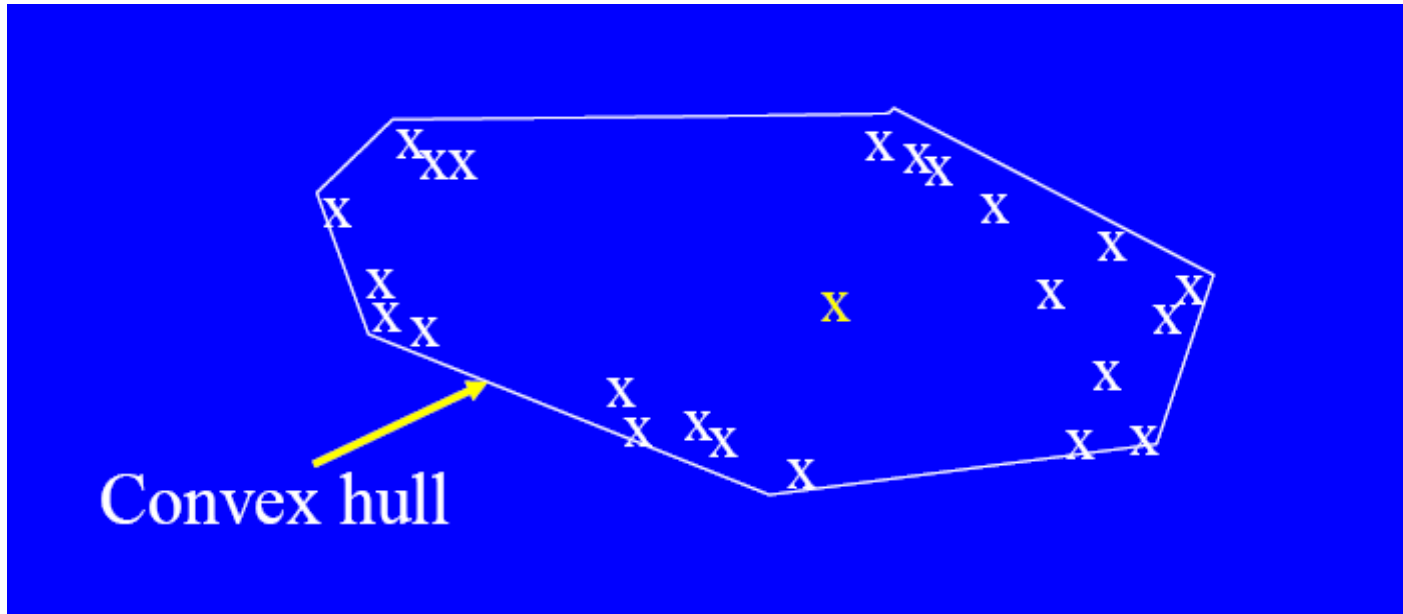
- ▶ Limitations

- ▶ Time consuming
- ▶ Subjective



Convex Hull Method

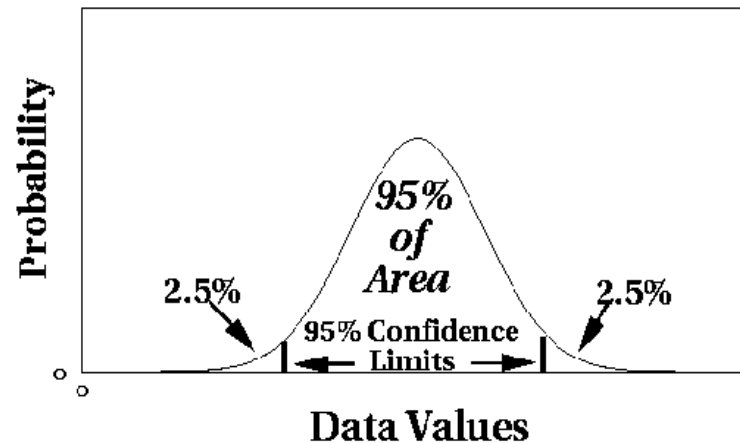
- ▶ Extreme points are assumed to be outliers
- ▶ Use convex hull method to detect extreme values



- ▶ What if the outlier occurs in the middle of the data?

Statistical Approaches

- ▶ Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- ▶ Apply a statistical test that depends on
 - ▶ Data distribution
 - ▶ Parameter of distribution (e.g., mean, variance)
 - ▶ Number of expected outliers (confidence limit)



Grubbs' Test

- ▶ Detect outliers in univariate data
- ▶ Assume data comes from normal distribution
- ▶ Detects one outlier at a time, remove the outlier, and repeat
 - ▶ H_0 : There is no outlier in data
 - ▶ H_A : There is at least one outlier

- ▶ Grubbs' test statistic:
$$G = \frac{\max |X - \bar{X}|}{s}$$

- ▶ Reject H_0 if:

$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N, N-2)}}{N-2 + t^2_{(\alpha/N, N-2)}}}$$



Statistical-based – Likelihood Approach

- ▶ Assume the data set D contains samples from a mixture of two probability distributions:
 - ▶ M (majority distribution)
 - ▶ A (anomalous distribution)
- ▶ General Approach:
 - ▶ Initially, assume all the data points belong to M
 - ▶ Let $L_t(D)$ be the log likelihood of D at time t
 - ▶ For each point x_t that belongs to M , move it to A
 - ▶ Let $L_{t+1}(D)$ be the new log likelihood.
 - ▶ Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
 - ▶ If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A



Limitations

- ▶ Most of the tests are for a single attribute
- ▶ In many cases, data distribution may not be known
- ▶ For multi-dimensional data, it may be difficult to estimate the true distribution



Distance-based Approaches

- ▶ Data is represented as a vector of features
- ▶ Three major approaches
 - ▶ Nearest-neighbor based
 - ▶ Density based
 - ▶ Clustering based



Nearest-Neighbor Based Approach

- ▶ Approach:
 - ▶ Compute the distance between every pair of data points
 - ▶ There are various ways to define outliers:
 - ▶ Data points for which there are fewer than p neighboring points within a distance D
 - ▶ The top n data points whose distance to the k th nearest neighbor is greatest
 - ▶ The top n data points whose average distance to the k nearest neighbors is greatest



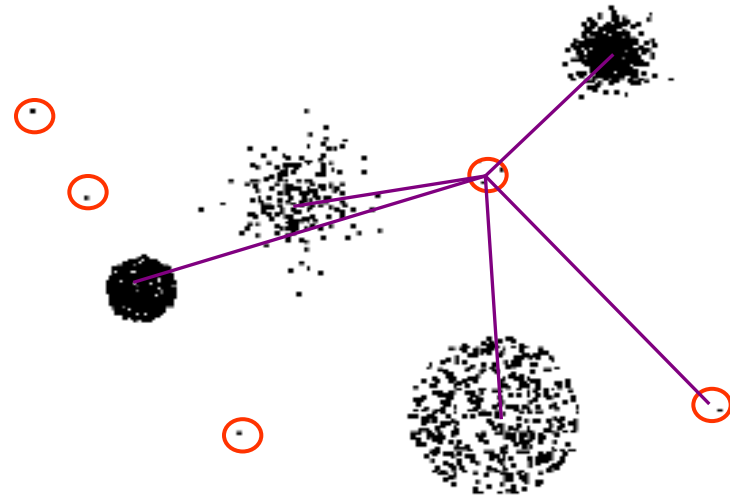
Density-based: LOF approach

- ▶ For each point, compute the density of its local neighborhood
- ▶ Compute local outlier factor (LOF) of a sample p as the average of the ratios of the density of sample p and the density of its nearest neighbors
- ▶ Outliers are points with largest LOF value



Clustering-Based

- ▶ Basic idea:
 - ▶ Cluster the data into groups of different density
 - ▶ Choose points in small cluster as candidate outliers
 - ▶ Compute the distance between candidate points and non-candidate clusters.
 - ▶ If candidate points are far from all other non-candidate points, they are outliers



Outliers in Lower Dimensional Projections

- ▶ In high-dimensional space, data is sparse and notion of proximity becomes meaningless
 - ▶ Every point is an almost equally good outlier from the perspective of proximity-based definitions
- ▶ Lower-dimensional projection methods
 - ▶ A point is an outlier if in some lower dimensional projection, it is present in a local region of abnormally low density



-
- ▶ Divide each attribute into ϕ equal-depth intervals
 - ▶ Each interval contains a fraction $f = 1/\phi$ of the records
 - ▶ Consider a d-dimensional cube created by picking grid ranges from d different dimensions
 - ▶ If attributes are independent, we expect region to contain a fraction f^k of the records
 - ▶ If there are N points, we can measure sparsity of a cube D as:

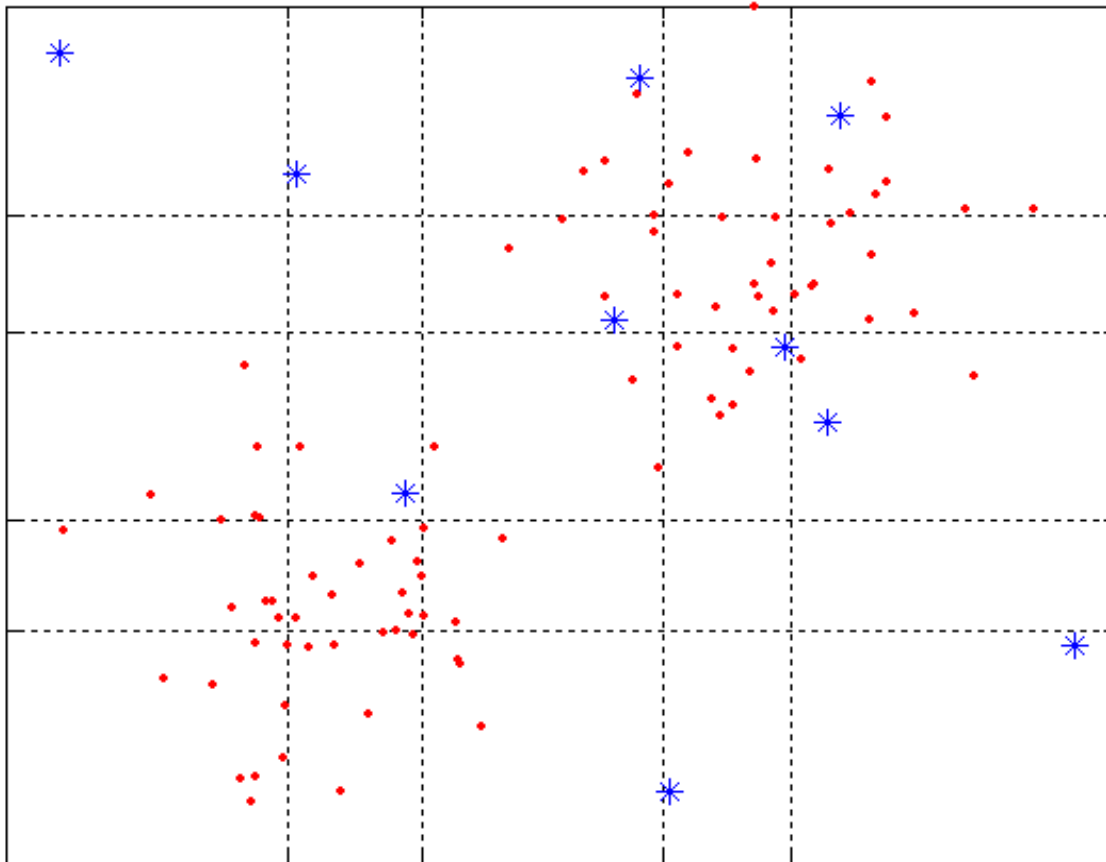
$$S(\mathcal{D}) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$

- ▶ Negative sparsity indicates cube contains smaller number of points than expected
 - ▶ To detect the sparse cells, you have to consider all cells.... exponential to d. Heuristics can be used to find them...
-



Example

- ▶ $N=100, \phi = 5, f = 1/5 = 0.2, N \times f^2 = 4$



How to treat outliers?

- ▶ **Trimming:** It excludes the outlier values from our analysis. By applying this technique our data becomes thin when there are more outliers present in the dataset. Its main advantage is its **fastest** nature.
- ▶ **Capping:** In this technique, we cap our outliers data and make the limit i.e, above a particular value or less than that value, all the values will be considered as outliers, and the number of outliers in the dataset gives that capping number.



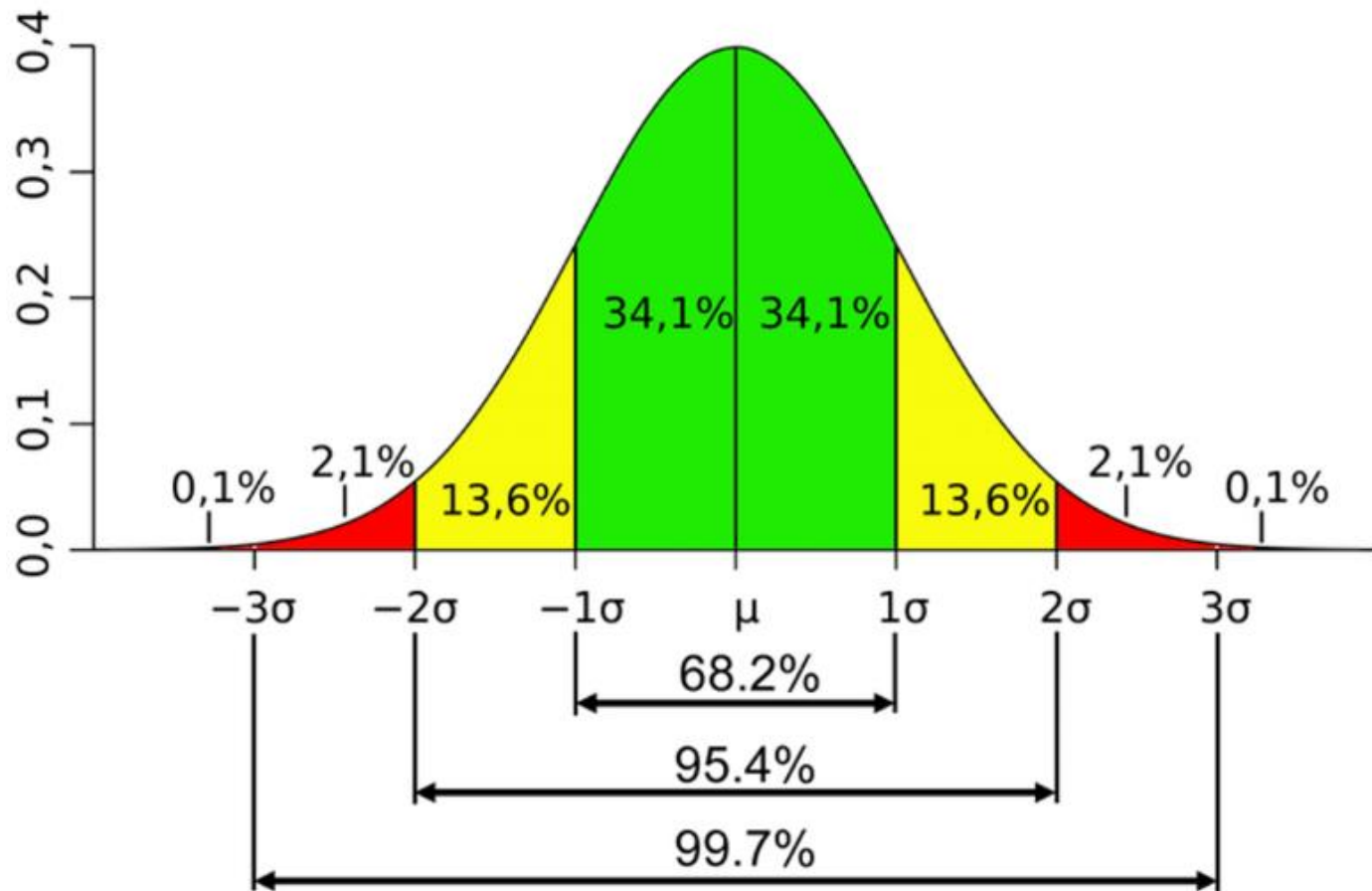
-
- ▶ **For Example**, if you're working on the income feature, you might find that people above a certain income level behave in the same way as those with a lower income. In this case, you can cap the income value at a level that keeps that intact and accordingly treat the outliers.
 - ▶ **Treat outliers as a missing value:** By assuming outliers as the missing observations, treat them accordingly i.e, same as those of missing values.
 - ▶ **Discretization:** In this technique, by making the groups, include the outliers in a particular group and force them to behave in the same manner as those of other points in that group. This technique is also known as **Binning**.
-



How to detect outliers?

- ▶ **For Normal distributions:** Use empirical relations of Normal distribution.
- ▶ – The data points which fall below ***mean-3*(sigma)*** or above ***mean+3*(sigma)*** are outliers.
- ▶ where mean and sigma are the **average value** and **standard deviation** of a particular column.



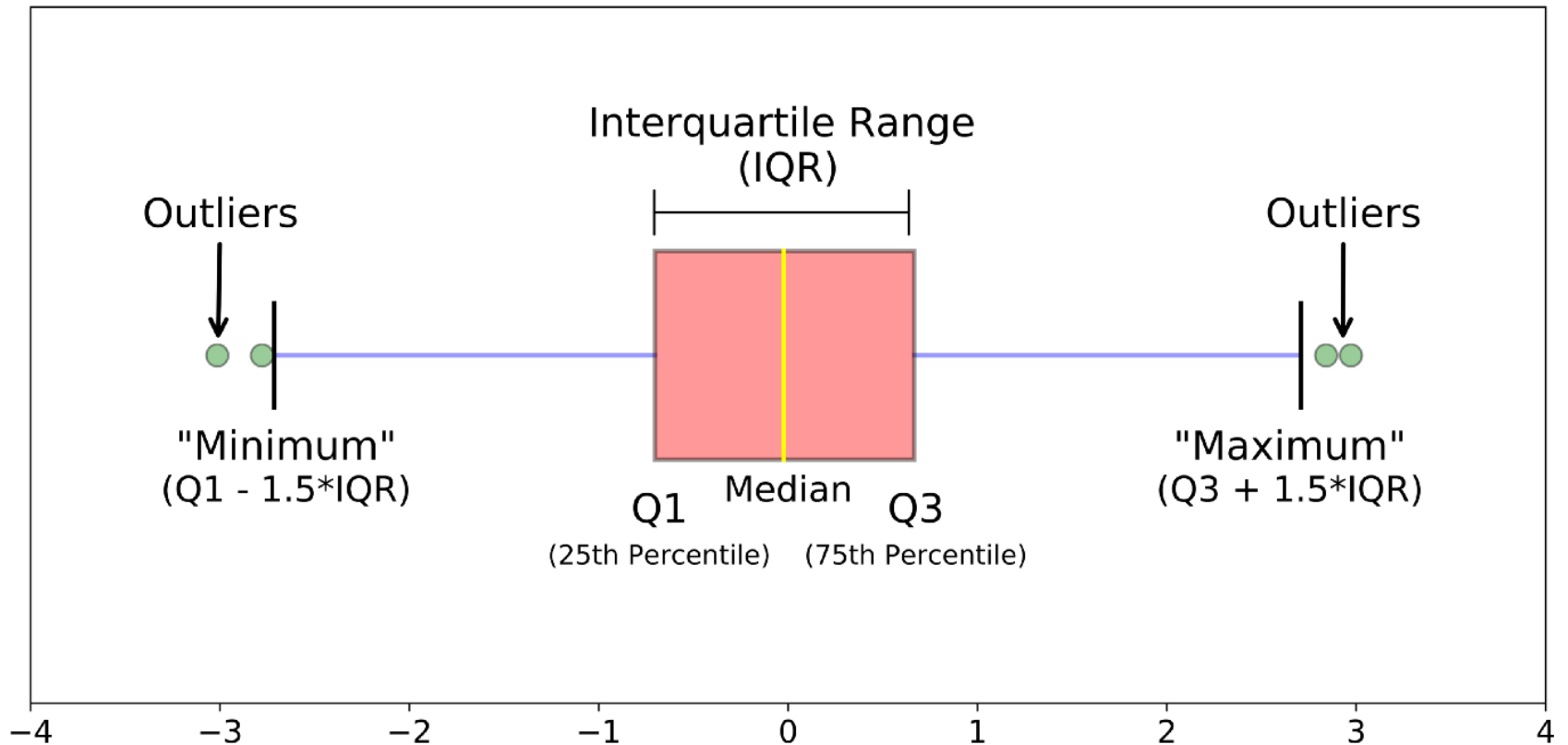


Z:

For Skewed distributions

- ▶ Use Inter-Quartile Range (IQR) proximity rule.
- ▶ – The data points which fall below **$Q1 - 1.5 IQR$** or above **$Q3 + 1.5 IQR$** are outliers.
- ▶ where $Q1$ and $Q3$ are the **25th** and **75th percentile** of the dataset respectively, and IQR represents the inter-quartile range and given by $Q3 - Q1$.





-
- ▶ **For Other distributions:** Use percentile-based approach.
 - ▶ **For Example,** Data points that are far from 99% percentile and less than 1 percentile are considered an outlier.



Techniques for outlier detection and removal

- ▶ **Z-score treatment :**
- ▶ **Assumption**– The features are normally or approximately normally distributed.













Deviation based Outlier detection

- ▶ Identifies outliers by examining the main characteristics of objects in a group
- ▶ Objects that “deviate” from this description are considered outliers
- ▶ Sequential exception technique
 - ▶ Simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects



Distribution and Plots

- ▶ A sample of data will form a distribution, and by far the most well-known distribution is the Gaussian distribution, often called the Normal distribution.
 - ▶ The distribution provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space.
 - ▶ This distribution describes the grouping or the density of the observations, called the probability density function.
 - ▶ also calculate the likelihood of an observation having a value equal to or lesser than a given value.
 - ▶ A summary of these relationships between observations is called a cumulative density function.
-



Distributions

- ▶ The distribution of a statistical dataset is the spread of the data which shows all possible values or intervals of the data and how they occur.
- ▶ A distribution is simply a collection of data or scores on a variable. Usually, these scores are arranged in order from ascending to descending and then they can be presented graphically.
- ▶ The distribution provides a parameterized mathematical function which will calculate the probability of any individual observation from the sample space.



-
- ▶ Data is a collection of information (numbers, words, measurements, observations) about facts, figures and statistics collected together for analysis.
 - ▶ **Example:** Distribution of **Categorical** Data (True/False, Yes/No): It shows the number (or) percentage of individuals in each group.
 - ▶ How to **Visualize** Categorical Data: Bar Plot, Pie Chart and Pareto Chart.
 - ▶ Distribution of **Numerical** Data (Height, Weight and Salary): Firstly, it is sorted from ascending to descending order and grouped based on similarity. It is represented in graphs and charts to examine the amount of variance in the data.
 - ▶ How to **Visualize** Numerical Data: Histogram, Line Plot and Scatter Plot.
-



Why are distributions important?

- ▶ Sampling distributions are important for statistics because need to collect the sample and estimate the parameters of the population distribution.
- ▶ Hence distribution is necessary to make inferences about the overall population



Difference between Frequency and Probability Distribution

S.No	Frequency Distribution	Probability Distribution
1	It records how often an event occurs. It is based on actual observations $\hat{=}$	It records the likelihood that an event is to occur. It is based on theoretical assumption of what should happen



Frequency Distribution:

- ▶ The number of times each numerical value occurs.

Example: Goals

Sam's team has scored the following numbers of goals in recent games

2, 3, 1, 2, 1, 3, 2, 3, 4, 5, 4, 2, 2, 3

Sam put the numbers in order, then added up:

- how often 1 occurs (2 times),
- how often 2 occurs (5 times),
- etc,

and wrote them down as a **Frequency Distribution** table.

From the table we can see interesting things such as

- getting 2 goals happens most often
- only once did they get 5 goals

Scores:
1,1,2,2,2,2,2,3,3,3,3,4,4,5

Score	Frequency
1	2
2	5
3	4
4	2
5	1

Probability Distribution

Consider tossing a fair coin 3 times.
Define X = the number of heads obtained

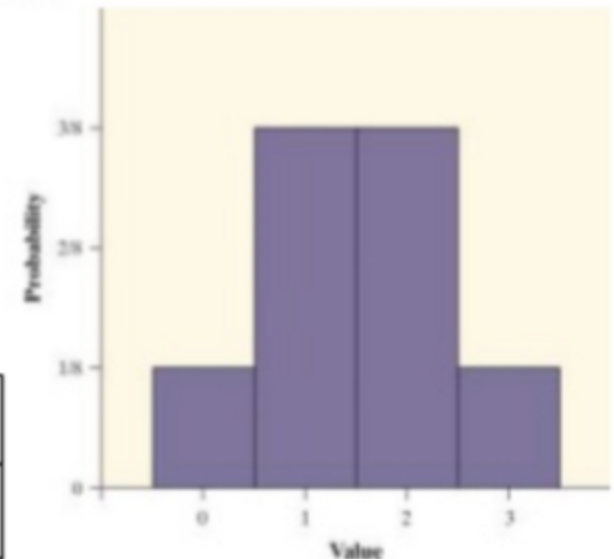
$X = 0$: TTT

$X = 1$: HTT THT TTH

$X = 2$: HHT HTH THH

$X = 3$: HHH

Value	0	1	2	3
Probability	$1/8$	$3/8$	$3/8$	$1/8$



Types of Distributions

- ▶ Bernoulli Distribution
- ▶ Uniform Distribution
- ▶ Binomial Distribution
- ▶ Normal Distribution
- ▶ Poisson Distribution
- ▶ Exponential Distribution



Bernoulli Distribution

- ▶ A special case of binomial distribution. It is the discrete probability distribution and has exactly only two possible outcomes – 1 (Success) and 0 (Failure) and a single trial.
- ▶ **Example:** In Cricket: Toss a Coin leads to win or lose the toss. There is no intermediate result. The occurrence of a head denotes success, and the occurrence of a tail denotes failure.
- ▶ The probability of success (1) is 0.4 and failure(0) is 0.6

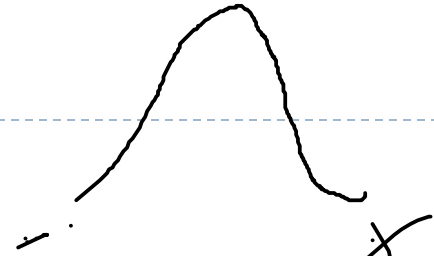


Normal Distribution

- ▶ It is otherwise known as Gaussian Distribution and Symmetric Distribution. It is a type of continuous probability distribution which is symmetric to the mean. The majority of the observations cluster around the central peak point.



- ▶ It is a bell-shaped curve.
- ▶ **Examples:** Performance appraisal, Height, BP, measurement error and IQ scores follow a normal distribution.
- ▶ *Mean = Median = Mode*



- ▶ **Basic Properties:**

- ▶ The normal distribution always run between $-\alpha$ and $+\alpha$
- ▶ Zero skewness and distribution is symmetrical about the mean.
- ▶ Zero kurtosis
- ▶ 68% of the values are within 1 SD of the mean
- ▶ 95% of the values are within 2 SD of the mean
- ▶ 99.7% of the values are within 3 SD of the mean

Binomial Distribution

- ▶ The most widely known discrete probability distribution. It has been used hundreds of years.

Binomial Distribution Formula

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

... 1 1

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial



▶ **Assumptions:**

- ▶ The experiment involves n identical trials. | -
- ▶ Each trial has only two possible outcomes – success or failure.
- ▶ Each trial is independent of the previous trials.
- ▶ The terms p and q remain constant throughout the experiment, where p is the probability of getting a success on any one trial and $q = (1 - p)$ is the probability of getting a failure on any one trial.



Poisson Distribution

- ▶ It is the discrete probability distribution of the number of times an event is likely to occur within a specified period of time. It is used for independent events which occur at a constant rate within a given interval of time.
- ▶ The occurrences in each interval can range from zero to infinity (0 to ∞).

2.01pm to 2.30pm
1.



► **Examples:**

- How many black colours are there in a random sample of 50 cars
- No of cars arriving at a car wash during a 20 minute time interval

Poisson Distribution Formula

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

$x = 0, 1, 2, 3, \dots$

λ = mean number of occurrences in the interval

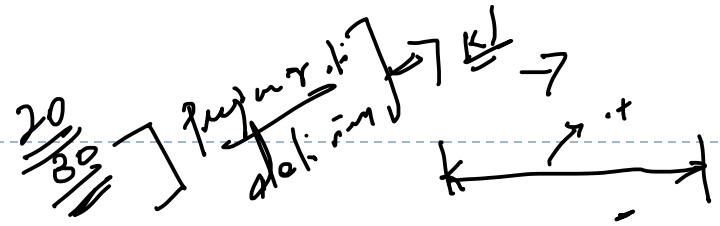
e = Euler's constant ≈ 2.71828



Uniform Distribution

- ▶ It is a continuous or rectangular distribution. It describes an experiment where an outcome lies between certain boundaries.





► **Examples:**

- Time to fly from Newark to Atlanta ranges from 120 to 150 minutes if we monitor the fly time for many commercial flights it will follow more or less the uniform distribution.
- The time taken for the students to finish a one hour test may range from 50 mins to 60 mins. An equal number of students complete over 5 minutes interval within this range – 50, 54, 56, 58 and 60. The finishing time of the test can be approximated by a uniform distribution.
- Time for Pizza delivery from Nanganallur to Alandur may range from 20 to 30 mins uniformly from the time delivery man leaves the Pizza Hut.



Gamma Distribution

- ▶ It deals with continuous variables which take on a wide range of values such as individual call times. Based on which we can model probabilities across any range of possible values using a gamma distribution function. First one is shape parameter (α) and the second one is scale parameter (β).

► **Examples:**

- ▶ The amount of rainfall accumulated in a reservoir.
- ▶ The size of loan defaulters and aggregation of insurance claims
- ▶ The flow of items through manufacturing and distribution processes
- ▶ The load on web servers *traffic*

7 traffic

8 supply

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526


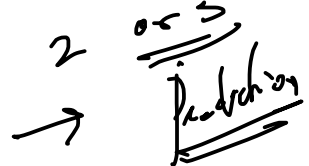



527

528

52

raise
down
 rance claims
 istribution
 7 Honor?
 7 day

Exponential Distribution

- ▶ It is concerned with the amount of time until some specific event occurs. 
- ▶ Example:
- ▶ The amount of time until an earthquake occurs has an exponential distribution 
- ▶ The amount of time in business telephone calls 
- ▶ The car battery lasts. 
- ▶ The amount of money customers spend on one trip to the supermarket follows an exponential distribution. There are more people who spend small amounts of money and fewer people who spend large amounts of money. 
- ▶ The exponential distribution is widely used in the field of reliability.









Density Functions

- ▶ Distributions are often described in terms of their density or density functions.
- ▶ Density functions are functions that describe how the proportion of data or likelihood of the proportion of observations change over the range of the distribution.
- ▶ Two types of density functions are probability density functions and cumulative density functions.
 - ▶ **Probability Density function:** calculates the probability of observing a given value.
 - ▶ **Cumulative Density function:** calculates the probability of an observation equal or less than a value.



Probability density function

- ▶ A probability density function, or PDF, can be used to calculate the likelihood of a given observation in a distribution.
- ▶ It can also be used to summarize the likelihood of observations across the distribution's sample space.
- ▶ Plots of the PDF show the familiar shape of a distribution, such as the bell-curve for the Gaussian distribution.



Cumulative density function

- ▶ A cumulative density function, or CDF, is a different way of thinking about the likelihood of observed values.
- ▶ Rather than calculating the likelihood of a given observation as with the PDF, the CDF calculates the cumulative likelihood for the observation and all prior observations in the sample space.
- ▶ It allows to quickly understand and comment on how much of the distribution lies before and after a given value.
- ▶ A CDF is often plotted as a curve from 0 to 1 for the distribution.



-
- ▶ Both PDFs and CDFs are continuous functions.
 - ▶ The equivalent of a PDF for a discrete distribution is called a probability mass function, or PMF.



Gaussian Distribution

- ▶ A Gaussian distribution can be described using two parameters:
 - ▶ **mean:** Denoted with the Greek lowercase letter μ , is the expected value of the distribution.
 - ▶ **variance:** Denoted with the Greek lowercase letter σ^2 (because the units of the variable are squared), describes the spread of observation from the mean.
- ▶ It is common to use a normalized calculation of the variance called the standard deviation
 - ▶ **standard deviation:** Denoted with the Greek lowercase letter σ , describes the normalized spread of observations from the mean.



Student's t-Distribution

- ▶ It is a distribution that arises when attempting to estimate the mean of a normal distribution with different sized samples.
- ▶ it is a helpful shortcut when describing uncertainty or error related to estimating population statistics for data drawn from Gaussian distributions when the size of the sample must be taken into account.
- ▶ **number of degrees of freedom:**
 - ▶ The number of degrees of freedom describes the number of pieces of information used to describe a population quantity.



Chi-Squared Distribution

- ▶ The chi-squared distribution is also used in statistical methods on data drawn from a Gaussian distribution to quantify the uncertainty.
- ▶ The chi-squared distribution has one parameter:
 - ▶ *degrees of freedom, denoted k .*
- ▶ An observation in a chi-squared distribution is calculated as the sum of k squared observations drawn from a Gaussian distribution.







