Quick Guide

Elements of Statistics and Probability (STA201)

KAZI SAKIB HASAN (24341237)

COMPUTER SCIENCE PROGRAM, BRAC UNIVERSITY

Acknowledgement

INTRODUCTORY STATISTICS – PREM S. MANN

INTRODUCTORY STATISTICS – BARBARA ILLOWSKY

PRACTICAL STATISTICS FOR DATA SCIENTISTS – PETER BRUCE

MD. SABBIR RAHMAN [SBBH], LECTURER (STA201), DEPARTMENT OF MNS, BRAC UNIVERSITY

Chapter 0: Pre-Requisite

- DISTRIBUTION AND VARIABLES
- RELATIVE FREQUENCY AND PROBABILITY
- GROUP AND UNGROUP DATA
- POPULATION VS SAMPLE
- PROBABILITY SAMPLING METHODS
- MEAN, VARIANCE AND STANDARD DEVIATION
- SAMPLING DISTRIBUTION, POINT ESTIMATE, AND CENTRAL LIMIT THEOREM
- NORMAL DISTRIBUTION AND THE EMPIRICAL RULE
- OUTLIER AND OUTLIER DETECTION USING EMPIRICAL RULE

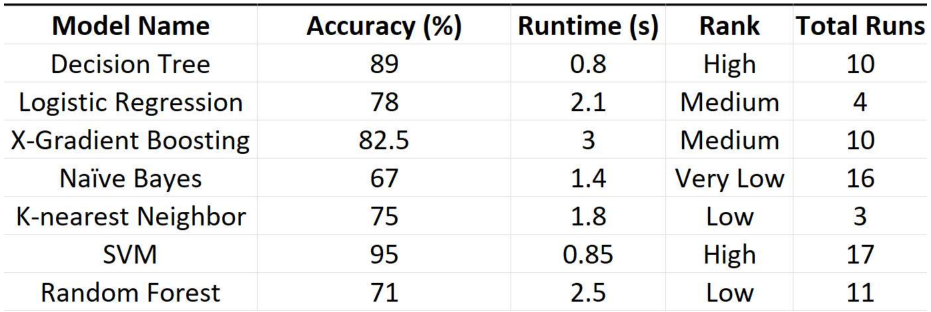
Distribution and Variables

- So, first of all, what is **distribution**?
- Distribution refers to place the values of a variable in a data table, or on a graph paper.

- Now, what is **variable**?
- A variable simply refers to the columns of a tabular dataset that take random values. For example, machine accuracy, bacteria life time, temperature, voltage amount are some examples of **variables**.

We will be using the words "**distribution**", "**normal distribution**", **"variables"**, and **"random variables"** a lot in this course.

Distribution and Variables



Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

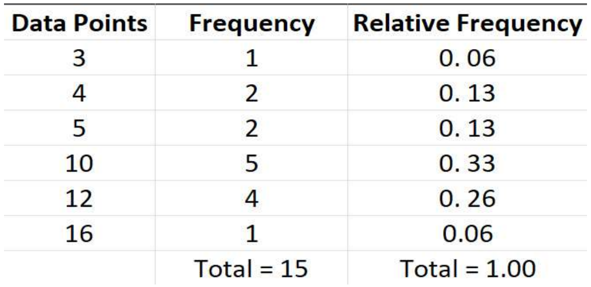
Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

In the dataset left to the screen, **Variables**: Model Name, Accuracy (%), Runtime (s), Rank, and Total Runs.

The quantitative variables (Accuracy, Runtime, and Total Runs) are distributed. We usually do not use the word "Distribution" for qualitative variables (Model Name, and Rank).

Relative Frequency and Probability

Example 01: Assume, a variable X consists of these datapoints: [3, 4, 4, 5, 5, 10, 10, 10, 10, 10, 12, 12, 12, 12, 16]. Create a frequency distribution table.



Data Points	Frequency	Relative Frequency
3	1	0. 06
4	2	0. 13
5	2	0. 13
10	5	0. 33
12	4	0. 26
16	1	0.06
	Total = 15	Total = 1.00

Synthetic Dataset 0.2: Frequency Distribution Table for X (Single-classed)

In statistics, distribution most commonly refers to the frequency count or relative frequency either of single datapoints or group data.

We use relative frequency to compute the probabilities. The probability of getting 10, 12, and 16 is respectively 0.33, 0.26 and 0.06. Confused?

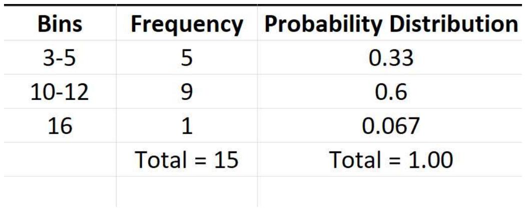
OK, assume X stands for the quiz scores. So, we can describe the dataset in the following manner –

1 student got 3.
2 students 4.
2 students got 5.
5 students got 10.
4 students got 12.
1 student got 16.
Total student = 1+2+2+5+4+1 = 15.

So, P(a randomly selected student got 10) = 5/15 = 0.33

Group Data

Example 02: Assume, a variable X consists of these datapoints: [3, 4, 4, 5, 5, 10, 10, 10, 10, 10, 12, 12, 12, 12, 16]. Create a frequency distribution table by grouping the data.



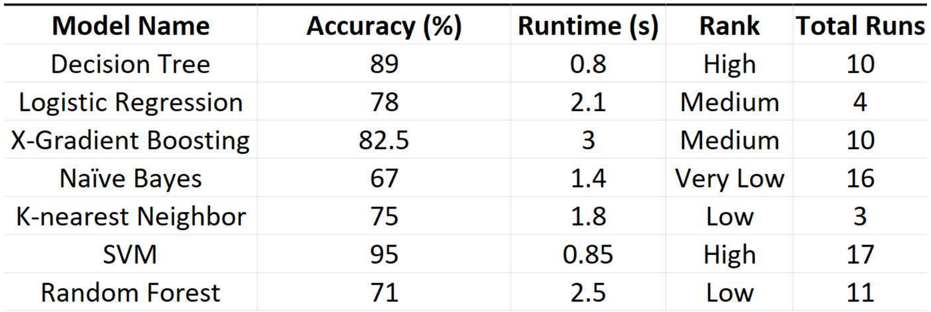
Bins	Frequency	Probability Distribution
3-5	5	0.33
10-12	9	0.6
16	1	0.067
	Total = 15	Total = 1.00

Synthetic Dataset 0.2: Frequency Distribution Table for X (Multi-classed/ Group data)

You can also represent data distribution by grouping them into specified class intervals / bins. Relative frequency and probability distribution stands for the same concept.

From the dataset, the probability of getting a number from 3 to 5 is 0.33
The probability of getting a number from 10 to 12 is 0.6
The probability of getting 16 is 0.067.

Distribution and Variables



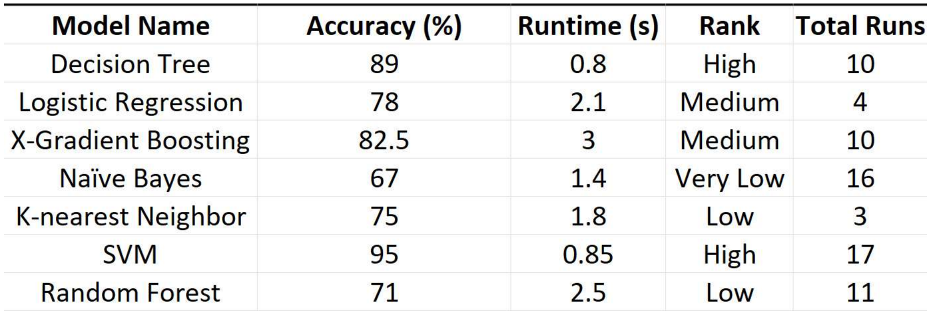
Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

In the dataset left to the screen, Variables: Model Name, Accuracy, Runtime, Rank, and Total Runs.

The quantitative variables (Accuracy, Runtime, and Total Runs) are distributed. We usually do not use the word "Distribution" for qualitative variables (Model Name, and Rank).

Variables



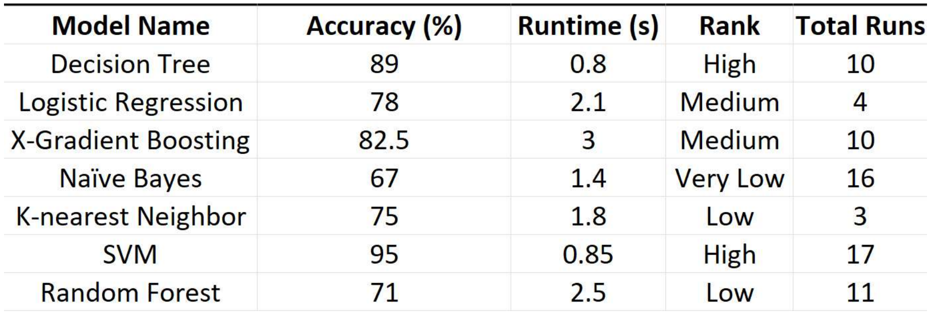
Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

Variables can be of two types.

- Qualitative / Categorical Variables
- Quantitative / Numerical Variables

Variables

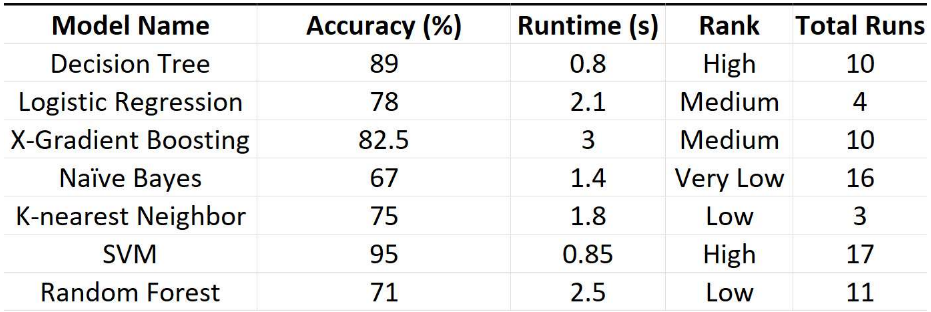


Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

Qualitative Variables: Variables that are non-numerical. In the left dataset, there are two qualitative variables which are "**Model Name**" and "**Rank**". These are also known as categorical variables.

Variables



Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

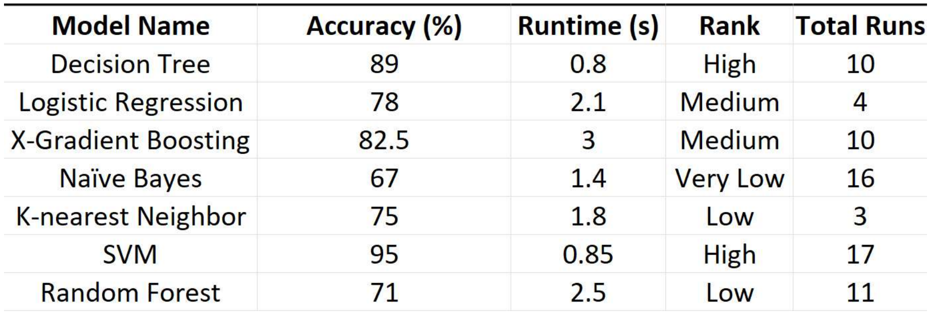
Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

Qualitative variables are made of – Nominal Data and Ordinal Data.

Nominal Data: Random non-numerical data that cannot be ordered (e.g. Model Name). It means, these data cannot be represented as finite sequence of unique numbers for practical use.

If you want to assign unique numbers to each of the model names (you don't know how many models there can be), there can be a lot of unique numbers (infinite) that you cannot even track. It will decrease computational efficiency for programming software.

Variables



Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

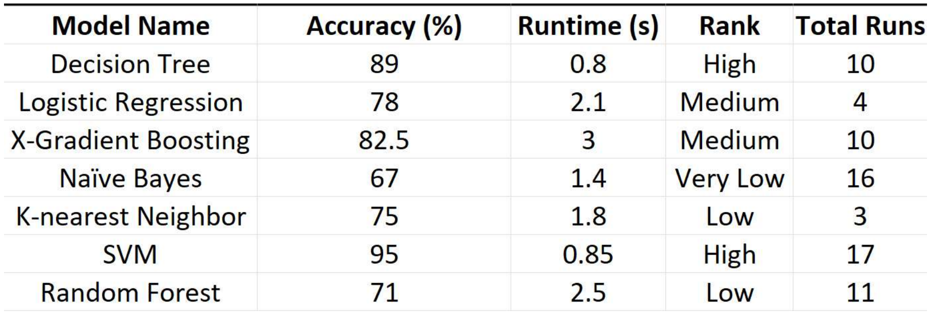
Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

Qualitative variables are made of – Nominal Data and Ordinal Data.

Ordinal Data: Data that are non-numerical but can be represented as ascending ordered numbers for practical usage (e.g. Rank). We can represent these ranks as Very Low (0), Low (1), Medium (2), High (3).

By assigning numbers to ordinal string variables, we can increase computational efficiency of programming software.

Variables

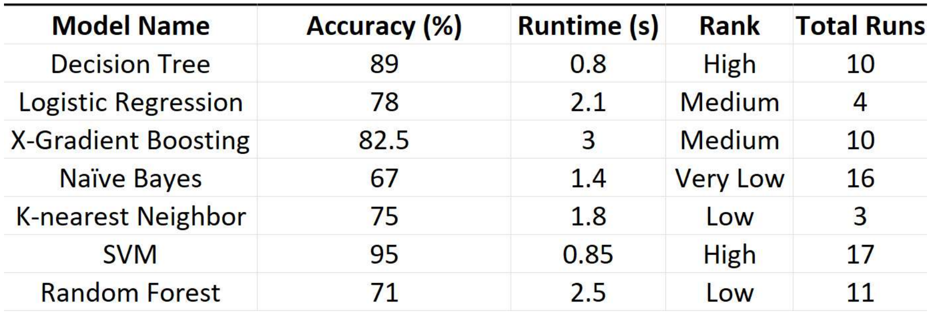


Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

Quantitative Variables: Variables that are numerical. In the left dataset, there are three quantitative variables which are **"Accuracy (%)"**, **"Runtime (s)"**, and **"Total Runs"**. These are also known as numerical variables.

Variables



Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

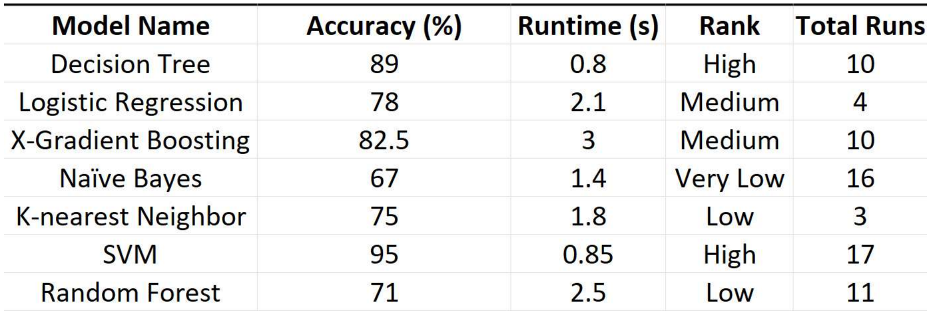
Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

Quantitative variables are formed with-
Discrete Data and Continuous Data

Discrete Data: Data that cannot take infinite numbers in between two positive consecutive numbers. For example, 1 and 2 are two consecutive positive numbers. Any numbers between 1 and 2 are not discrete data (e.g. 1.5, 1.6, 1.9 etc.)

So, you may say that discrete data are simply positive integers. This is true for almost every practical scenario. In our left dataset, **Total Runs** is a discrete quantitative variable. A user cannot run a model 10.4 or any other float number times.

Variables



Model Name	Accuracy (%)	Runtime (s)	Rank	Total Runs
Decision Tree	89	0.8	High	10
Logistic Regression	78	2.1	Medium	4
X-Gradient Boosting	82.5	3	Medium	10
Naïve Bayes	67	1.4	Very Low	16
K-nearest Neighbor	75	1.8	Low	3
SVM	95	0.85	High	17
Random Forest	71	2.5	Low	11

Synthetic Dataset 0.1: Supervised Machine Learning Models' efficiency record for a random Artificial Intelligence project.

Quantitative variables are formed with- Discrete Data and Continuous Data

Continuous Data: Data that can take infinite numbers in between two consecutive numbers. For example, 1 and 2 are two consecutive numbers. Any numbers from 1 to 2 including are continuous data (e.g. 1, 1.2, 1.3, 2 etc.). Negative integers or floats are also continuous data.

So, you may say that continuous data are simply integer and floats both. In our left dataset, **Accuracy (%)**, and **Runtime (s)** are continuous quantitative variables.

Population vs Sample

- The set of every potential element/participant for a specific research project is known as the **population.** The characteristics (e.g. mean, median, mode, sd etc.) of population is known as **parameter.**
- The subset of population is known as the **sample**. The characteristics of sample is known as **statistic**.

Example

Imagine you want to calculate the proportion of defective resistors in your EEE lab to decide whether you should stay or change the lab section for convenience.

- There can be total **N** resistors in your lab which denotes the **population**.
- You can select **n** resistors from the population as representative **sample** to estimate the proportion.

Population vs Sample

- The set of every potential element/participant for a specific research project is known as the **population.**
- The subset of population is known as the **sample**.

Example

Imagine you want to calculate the proportion of defective resistors in your EEE lab to decide whether you should stay or change the lab section for convenience.

- It will be really time consuming if you test every resistor in the lab. That's why you can take several resistors as sample from each box and calculate the proportion of defection. This will give you an estimation of how many resistors are defective in your lab section.

Population vs Sample

- **Physical Population :** The size of the population may be unknown but can be calculated somehow. Example: Set of every existing resistor in your university.
- **Conceptual Population:** The size of the population is unknown and cannot be calculated. Example: Resistors that are being manufactured in an industry.
- **Physical Sample:** Subset of Physical Population.
- **Conceptual Sample:** Subset of Conceptual Sample.

Population vs Sample

- The total resistor in your university can be calculated, it means that they **already** physically exists. There are already **N** numbered resistors in your university. That is why the total number of resistors in your university is an example of Physical Population.
- Conversely, imagine the resistors are manufactured in “X” industry by applying several method. To test the method, the manufacturers will create **n** numbered sample of resistors and then test them. This sample is a part of the population that yet doesn’t physically exist. So, it is a **conceptual sample**. Because the manufacturers are not sure yet how many resistors following the same method will be produced (**N**). That is why this is an example of **Conceptual Population**.

Probability Sampling Methods

- The sample you should take from the population must be **representative** (at least in this course). You can apply probability sampling techniques to ensure a representative sample. Every element in a population has equal probability of being selected as a sample element in probability sampling techniques. Similarly, in a **representative sample**, every sample element had the equal probability of being selected from the population.
- Two common probability sampling techniques –
- Simple Random Sampling (most used)
- Systematic Sampling

Probability Sampling Methods

- **Simple Random Sampling (SRS)**: Randomly pick sample elements from the population elements without any bias. We can use a random integer number generator for this.
- Imagine, N = 30 and we need to create a sample where n = 5. Write a computer program that will randomly generate five random integers within the interval 1 to 30. Assume that the numbers are 5, 10, 13, 23, 27. Index every population element from 1 to 30 and pick 5th, 10th, 13th, 23rd , and 27th population elements as sample elements.
- **Systematic Sampling**: Assign a value for *k* and pick every *k*th element from the population as a sample element.
- Imagine the previous scenario. Let, k = 6. So our sampling elements are going to be 1st, 6th, 12th, 18th, and 24th element from the population elements.

Mean

- Assume a variable X with the following data points:

X = [5, 1.25, 6, 3, 2.25, 6.75]

Mean = (5 + 1.25 + 6 + 3 + 2.25 + 6.75) / 6 = 4.04 (Ans)

Formula: Mean = summation of data points / number of data points

- Number of data points is often referred as sample size, data length, or population size based on the context.
- Mean is also called as arithmetic mean and average.
- Mean is the most common measure of central tendency and is used to find the central point/midpoint of a sorted or unsorted univariate dataset.

Dataset with one variable

Mean

- Assume a variable X that denotes the population data (N = 300). The population mean of X is defined by μ.

  In this context, population mean μ = $\frac{x_1+..+x_{300}}{300}$

- Now, assume a sample created with 50th to 150th elements of the population. The sample mean of X is defined by $\bar{x}$.

  In this context, sample mean $\bar{x}$ = $\frac{x_{50}+..+x_{150}}{(150-5\ \ )}$

Variance and Standard Deviation (Sample)

- Calculating variance and standard deviation for the same random variable X = [5, 1.25, 6, 3, 2.25, 6.75], assuming that X is a sample of a larger population.
- Sample Variance s^2 = $\frac{(4.04-5)^2+(4.04-1.25)^2+(4.04-6)^2+(4.04-3)^2+(4.04-2.25)^2+(4.04-6.75)^2}{6-1}$

  = 4.83 (Ans)

Sample Standard Deviation, s = $\sqrt{Sample\ Variance}$ = 2.19 (Ans)

Variance and Standard Deviation (Population)

- Calculating variance and standard deviation for the same random variable X = [5, 1.25, 6, 3, 2.25, 6.75], assuming that X is the population.
- Population Variance σ^2= $\frac{(4.04-5)^2+(4.04-1.25)^2+(4.04-6)^2+(4.04-\ \ )^2+(4.04-\ \ .25)^2+(4.04-6.75)^2}{6}$

  = 4.025 (Ans)

Population Standard Deviation, σ = $\sqrt{Sample\ Variance}$ = 2.006 (Ans)

- Standard deviation is the most common measure of dispersion. It tells how far the data points are from the mean. It therefore gives a clear overview on the fluctuation of data points. A machine is considered to be more efficient if it has relatively less standard deviation in efficiency/accuracy than the other machines. However, the interpretation of standard deviation should be done cautiously based on the context.

Visualizing Standard Deviation

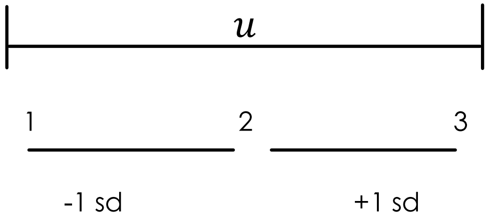
- Assume, a variable X = $[x_1, x_2, x_3]$, where $x_1 = 1, x_2 = 2, x_3 = 3$

Hence, X = [1,2,3]

Now,

Mean of X = 2

Sample Standard deviation of X = 1



u

1 2 3

-1 sd +1 sd

Notice that, x_1 and x_3 both are 1 standard deviation far from the mean (x_2).

Mean – standard deviation = 2 – 1 = 1 (x_1)

Mean + standard deviation = 2 + 1 = 3 (x_3)

Sampling Distribution

- Imagine a variable Y of size **N** that represents population data. We can create $\binom{N}{n}$ numbered samples from the population, where n is the number of every sample size.
- Assume, population size N = 10. We want to know how many samples of size n = 5 can be created from the population.

Total possible samples = $\binom{10}{5} = 252$ (Ans)

Sampling Distribution

- Create a sampling distribution for Y = [70, 78, 80, 80, 95]

Assuming, A = 70, B = 78, C = 80, D = 80, E = 95

So, total possible samples are: ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE

(Continue to next slide..)

Sampling Distribution

Sample Name	Values	Mean ($\bar{x}$)
ABC	[70, 78, 80]	76
ABD	[70, 78, 80]	76
ABE	[70, 78, 95]	81
ACD	[70, 80, 80]	76.67
ACE	[70, 80, 95]	81.67
ADE	[70, 80, 95]	81.67
BCD	[78, 80, 80]	79.33
BCE	[78, 80, 95]	84.33
BDE	[78, 80, 95]	84.33
CDE	[80, 80, 95]	85

Table 0.1: All possible samples and their means ($\bar{x}$)

Sampling Distribution

Mean ($\bar{x}$)	P($\bar{x}$)
76	0.2
76.67	0.1
79.33	0.1
81	0.1
81.67	0.2
84.33	0.2
85	0.1

Table 0.2: Sampling distribution of $\bar{x}$

Facts: The mean of Y is **80.6**

If you calculate the mean of the column Mean ($\bar{x}$) from table 0.1, the result will be **80.6**

If you multiply the values of Mean ($\bar{x}$) with P($\bar{x}$) in Table 0.2 and add them, the result will be also **80.6**

Sampling Distribution

Mean ($\bar{x}$)	P($\bar{x}$)	$\bar{x}$ P($\bar{x}$)
76	0.2	15.2
76.67	0.1	7.667
79.33	0.1	7.933
81	0.1	8.1
81.67	0.2	16.334
84.33	0.2	16.866
85	0.1	8.5

$$\sum xP(x) = E(x) = 80.6$$

Table 0.3: Calculating expected value E(x).

The expected value is nothing but the mean of our original Y data. So, the interpretation for **mean** and **expected value** is the same.

Point Estimate

- The **calculated statistic** (e.g. mean, median, mode, sd etc.) from a sample is known as **point estimate**.
- **Sampling error** = calculated statistic – population parameter
- **Non-sampling error** = calculated statistic after causing human made mistake – population parameter.

Example 01: Population data = [10, 6, 18, 12.5, 11] and sample = [6, 18, 12.5]. Calculate sampling error of mean as a point estimate.

Solution: Sample mean = 12.167

Population mean = 11.5

Sampling error = 12.167 – 11.5 = 0.667 (Ans)

Now, if we calculated the sample mean by making a mistake while doing the calculation (e.g. entering 28 instead of 18 during calculation) it would create a non-sampling error by deviating the sampling mean to 15.5.

Non-sampling error = 15.5 – 11.5 = 4

Central Limit Theorem (CLT)

- For a population with size N > 30, if every sample with size n ≥ 30 from the population is taken, then the sampling distribution will be approximately normal. The population data does not need to be normally distributed in this case. For example,
- Imagine a population dataset with 35 data points (N = 35). If we want to take only 30 sized (n = 30) samples from the population-

There can be total $\binom{35}{30} = 324632$ samples.

Now, as we said, all these 324632 samples have a size of 30 (n = 30).

According to De Moivre's CLT, if you calculate the sampling distribution of all these 324632 samples (as shown in Table 0.1 – Table 0.2) and plot the distributions in a graph, the plot curve will be approximately symmetrical. In other words, the sampling distribution will be approximately normal.

Normal Distribution

- The distribution of random continuous variables that looks like symmetric bell-shaped curve if every data point of that variable, or the probability distribution is placed on a graph paper.
- In a normally distributed dataset, mean = median = mode.
- The total area under the curve is 1.00 (because the total sum of probability distribution is 1).
- This distribution is used in –
- Calculating confidence interval (CI) for population mean and population proportion.
- Hypothesis Testing for one or two samples.
- Calculating the probability of getting a value staying within certain interval in a normally distributed dataset.
- Calculating the probability of getting a value less than the given value in a normally distributed dataset.

Inspiring Excellence

The Empirical Rule (68-95-99.7 Rule)

- CLT by-produced the empirical rule. According to the rule, in a normally distributed dataset:

* Approximately 68.44% data points falls within the interval (μ - 1. σ, μ + 1. σ)

* Approximately 95% data points falls withing the interval (μ - 2. σ, μ + 2. σ)

* Approximately 99.7% data point falls within the interval (μ - 3. σ, μ + 3. σ)

Slide to next page for an illustrated example.

The Empirical Rule (68-95-99.7 Rule)

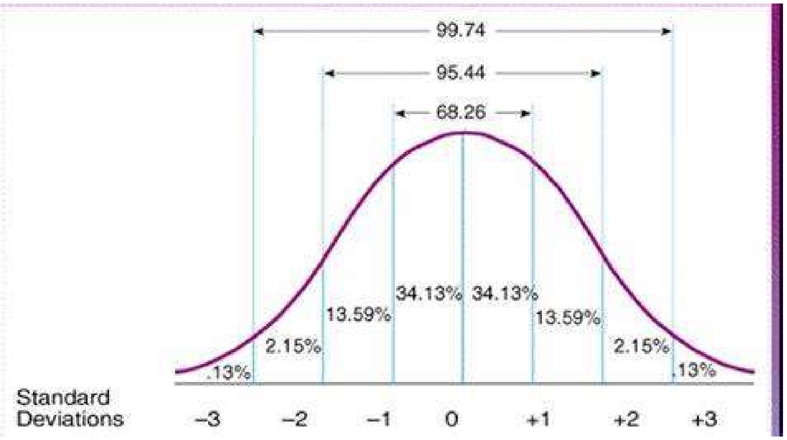


Fig 0.1: The normal distribution curve and the empirical rule

Facts: If you randomly take at least 30 numbers and make a plot, the plot will be similar to Fig 0.1. Assume, Fig 0.1 illustrates a normally distributed data consisting 100 data points. The mean and standard deviation of the data is approximately 0 and 1. According to the Empirical Rule,

68 data points will fall within (0-1, 0+1) interval.
95 data points will fall within (0 – 2*1, 0 + 2*1) interval.
99 data points will fall within (0 – 3*1, 0 + 3*1) interval.

The Empirical Rule (68-95-99.7 Rule)

- Normal Distribution is also known as Gaussian Distribution.
- The Central Limit Theorem was originally developed by Abraham De Moivre. Gauss formulated the normal distribution almost after 100 years of Central Limit Theorem.
- The symmetrical behavior of univariate dataset was first observed by Galileo when he was experimenting with astronomical errors during 17th century.

Outlier

- Datasets often have values/data points that are extremely larger or smaller than the usual values. These are known as **outliers** or **extreme values**.
- Assume, you are trying to point estimate the mean monthly pocket money of teenage girls in a city. You collected a sample of 50 girls and found out that 47 girls have their pocket money in the range 500 to 700 taka, except the three girls –
- One girl gets 0 taka.
- The second one gets 3000 Taka.
- The third one gets 4000 Taka.

Here, 0, 3000, 4000 are the outliers in the dataset. Your point estimate will be unrealistic if you do not cancel out these outliers. The small outliers (0 Taka) are known as **low outliers** and the large outliers (3000 and 4000 Taka) are known as **high outliers**.

Outlier Detection

Already mentioned that 99% data values fall within 3 standard deviation from the mean. What about the rest 1% data then? These data are usually outliers. Imagine a dataset with mean 10.5 and standard deviation of 1.25. Then,

- Values less than (10.5 – 3 * 1.25) = 6.75 are considered as **low outliers.**
- Values more than (10.5 + 3 * 1.25) = 14.25 are considered as **high outliers.**

However, this is not the usual rule to detect outliers. We usually use John Tukey's IQR Rule to detect outliers that we will discuss in the next chapter.

Chapter 01: Descriptive Statistics

KAZI SAKIB HASAN (24341237)

COMPUTER SCIENCE PROGRAM, BRAC UNIVERSITY

kazi.sakib.hasan@g.bracu.ac.bd

Brief Contents

- MEASURES OF CENTRAL TENDENCY
- QUANTILES
- MEASURES OF DISPERSION
- OUTLIER DETECTION (IQR RULE)
- BOXPLOT
- SKEWNESS
- STEM AND LEAF PLOT
- APPENDICES

Measures of Central Tendency

- Topics to discuss –
- Arithmetic mean, Weighted mean, Geometric mean, Harmonic mean.
- Median
- Mode

Note: We already discussed arithmetic mean in the pre-requisite chapter. So, we will skip it here.

Weighted Mean

- We calculate weighted mean when each value in a dataset have different weights. For arithmetic mean, the weight is constant, not different. Every value carries a weight of 1 in such datasets where we use arithmetic mean.
- Formula: Weighted mean = $\frac{x_1.w_1+x_2.w_2+..+x_n.w_n}{n}$

Weighted Mean

- **Example 01**: The importance of each feature (x_i) in a layer on neural networks' output is denoted by their respective weights. Assume a small layer where the features are 0.5, 0.3, 0.2 and the corresponding weights are 0.8, 0.4, and 0.6. Calculate the weighted mean.
- **Solution**: Weighted mean = $\frac{0.5*0.8+0.3*0.4+0.2*0.6}{3}$ = 0.213 (Ans)

Weighted Mean

- **Example 02**: There are 3 apples that weigh 100 gm, 5 apples weigh 150 gm and 7 apples weigh 125 gm. If a 100 gm apple contains 12 mg of Vitamin C, can we expect to get 80 mg of Vitamin C on average if we juice the apples?
- **Solution**: Weighted mean = $\frac{3*100+5*150+7*125}{3} = 641.66 \text{ gm}$

Now, 100 gm contains 12 mg Vitamin C.

So, 1 gm contains 12/100 = 0.12 mg vitamin C.

Then, 641.66 gm contains 0.12 * 641.66 = 76.99 mg vitamin C.

Therefore, we cannot expect to get 80 mg Vitamin C from the juice of these apples. (Ans)

Geometric Mean

- We use geometric mean when we deal with data regarding rates of change, growth factors or the data is log-normally distributed.
- It is usually used in :
- Business and Finance : Stock indexes, portfolio returns etc.
- Biology : Microbiology, biodiversity, epidemiology, pharmacokinetics etc.
- Environmental Science : Concentration of pollutants
- Physics : Frequencies of harmonic oscillations.
- Astronomy : Analysis of log-normal distributions of stellar properties.
- Formula: Geometric mean = $(x_1 * x_2 * x_3 * \ldots * x_n)^{1/n}$

Geometric Mean

- What is log-normal distribution?

- If you apply **exponential function** to every data point of a **normally distributed dataset**, then the new data points become log-normally distributed.

The data becomes **rightly skewed**.

Assume, a normally distributed dataset = [1, 1.2, 3, 3.5, 3.25, 3.75, 6]

Log-normal transformation = $[e^1, e^{1.2}, e^3, e^{3.5}, e^{3.25}, e^{3.75}, e^6]$

Here, e = Euler's number (2.718)

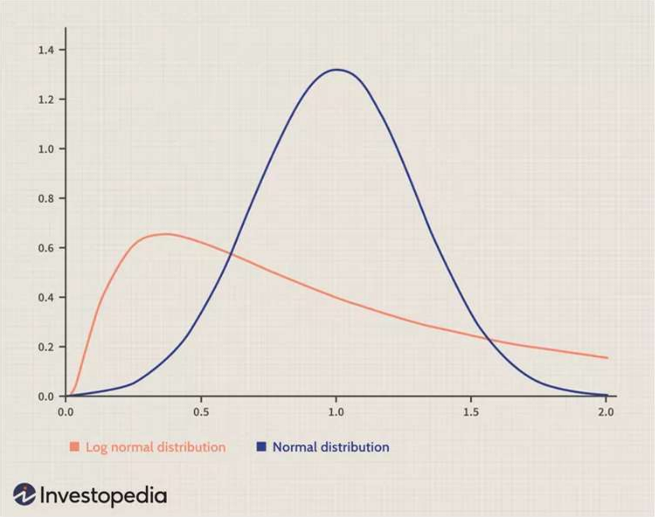


Fig 1.1: Curve difference between normally distributed dataset and its log-normal distribution.

Geometric Mean

Example 01: The bacterial concentrations in four water samples are 100 CFU/ml, 1000 CFU/ml, 10000 CFU/ml, 100000 CFU/ml. What is the average bacterial concentration?

Solution: Geometric mean of bacterial concentrations,

$(100 * 1000 * 10000 * 100000)^{1/4}$= 3162. 27 CFU/ml (Ans)

Why we did not use arithmetic mean? Because the data is **highly skewed** and have **exponential growth** and hence, arithmetic mean would produce bias result.

Harmonic Mean

- We use harmonic mean when the data points are **rates** or **ratios.**
- Most appropriate to use when data involves quantities like speed, densities, etc.
- Formula: Harmonic mean = $\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\frac{1}{x_4}+\cdots+\frac{1}{x_n}}$

Harmonic Mean

- **Example 01**: Suppose a car travels a certain distance at 60 km/h and returns over the same distance at 40 km/h. What is the average speed for the entire journey?

Solution (Physics approach): $t_1 = \frac{d}{60}$ and $t_2 = \frac{d}{40}$

Total time, $t = \frac{d}{60} + \frac{d}{40} = \frac{d}{24}$

Total Distance, $s = 2d$

So, average speed, $v = \frac{s}{t} = \frac{2d}{\frac{d}{24}} = 48\ km/h$ (Ans)

Harmonic Mean

- **Example 01:** Suppose a car travels a certain distance at 60 km/h and returns over the same distance at 40 km/h. What is the average speed for the entire journey?

Solution (Harmonic mean approach):

Average speed, $v = \frac{2}{\frac{1}{60}+\frac{1}{40}} = 48\ km/h$ (Ans)

Harmonic Mean

- **Example 02**: Consider a scenario where data packets are transmitted over three different network paths with latencies of 100 ms, 150 ms, and 300 ms. What is the average latency if the data is distributed evenly across these paths?

Solution: We will use harmonic mean instead of other means since the data points include **rate**.

$$H = \frac{3}{\frac{1}{100}+\frac{1}{150}+\frac{1}{300}} = 150\,ms \text{ (Ans)}$$

Median

- The **middle** data point of an odd number sized **sorted** univariate dataset is called as median. For example, assume a dataset X = [3, 5, 1, 8, 6] is. Here, dataset size, n = 5, which is an odd number. The sorted dataset is [1, 3, **5**, 6, 8]. Hence, the median is **5**.
- The **average of the two middle data points** of an even number sized **sorted** univariate dataset is called as median. For example, assume a dataset X = [3, 5, 1, 8, 7.5, 2]. Here, dataset size, n = 6, which is an even number. The sorted dataset is [1, 2, **3, 5**, 7.5, 8]. So, median is (3 + 5)/2 = **4.**

- For calculating median, you can sort the dataset in ascending order, or in descending order. The choice is yours.

Median

- **Example 01:** Calculate the median for the dataset Y = [14, 9, 13.75, 15, 3.5, 6.5, 14, 10, 13, 12.25]

Solution: Here, n = 10 (even number)

Step 01: So, median is the average of the two middle data points in sorted Y dataset.

Sorted_Y = [3.5, 6.5, 9, 10, 12.25, 13, 13.75, 14, 14, 15]

Step 02: The two middle data points are in $\frac{n}{2} th\ and\ \left(\frac{n}{2}+1\right)\mathrm{th}$ position in sorted Y dataset. Now, n = 10. So, $\frac{n}{2}$ and $\left(\frac{n}{2}+1\right)$ is respectively 5 and 6.

Step 03: So, the median is average of 5th and 6th element in sorted Y dataset. 5th and 6th element are respectively 12.25 and 13.

Median = $\frac{12.25+13}{2} = 12.625\ (Ans)$

Median

- **Example 02:** Calculate the median for the dataset Y = [14, 9, 13.75, 3.5, 6.5, 14, 10, 13, 12.25]

Solution: Here, n = 9 (odd number)

Step 01: So, median is the middle data point in sorted Y dataset.

Sorted_Y = [3.5, 6.5, 9, 10, 12.25, 13, 13.75, 14, 14]

Step 02: The middle data point is {(n//2) + 1}th data point. Here, 9//2 = 4, and 4+1 = 5. So, 5th data point in the sorted Y dataset is the median.

Step 03: The median is 12.25 (Ans).

- n//k means what? When you divide n by k where k is an odd number, you get a decimal number which is n/k. If you remove that decimal portion and keep the integer only, that is the resultant for n//k. For example, 9/2 = 4.5, but 9//2 = 4.

Median

- **Why do use median as a measure of central tendency?**
- Median is a stable and reliable measure due to its robustness to outliers. Every mean is affected by the outliers but medians are not. Because median is just the middle point in that particular sorted dataset, whereas the outliers are data points in the beginning or ending positions in that sorted dataset.
- When the data is skewed, (not normally distributed), means can be biased to the skewed side. But as median is the center point of that sorted dataset, it estimates the central tendency without any bias.
- Median is nothing but the data point that divides the entire dataset into two equal halves. The data points in the left portion of the median are smaller values than the median, and the right portion points are higher than the median.

Mode

- Mode is the data point that has maximum frequency (occurrence) in a univariate dataset.
- If there is only one mode in a dataset, then it is an unimodal dataset. For example, [1,1,1,3,4] is an unimodal dataset where the mode is 1.
- If there is two mode in a dataset, then it is an bimodal dataset. For example, [1,1,1,6,6,6,2,2] is an bimodal dataset where the mode is 1 and 6. Both 1 and 6 appeared 3 times.
- If there is more than two modes in a dataset, then it is an multimodal dataset. For example, [1,1,2,2,3,3,4,4,7,8,9] is a multimodal dataset where 1, 2, 3, and 4 are modes. They all appeared 2 times.

Mode

- **Why do we use mode as a measure of central tendency?**
- When the data type is nominal, we cannot use mean and median. In such case, we calculate mode. For example, the most preferred choice in a survey is calculated by mode.
- Mode indicates the most common data point in a dataset. It has several practical usages like most popular choice, observation, measurement etc.
- Similarly to median, modes are not affected by outliers.
- Mode tells us in which point in the axis, the distribution curve has its peak.

Quantiles

- Topics to discuss –

- Quartiles
- Deciles
- Percentiles

Quartiles, Deciles, and Percentiles

- **Quartiles:** Quartiles are the three data points, or a point within the univariate dataset that separates the dataset in four equal parts.

- First Quartile (Q1): The **25th** percentile, which separates the **lowest 25%** of the data from the rest.
- Second Quartile (Q2): The **50th** percentile, also known as the **median**, which divides the data into two equal halves.
- Third Quartile (Q3): The **75th** percentile, which separates the **lowest 75%** of the data from the highest 25%.

Quartiles, Deciles, and Percentiles

- **Deciles:** Similar to quartiles, deciles separate the dataset into 10 parts. These are termed as D1, D2, D3,….., D10. Each decile contains 10% of the data.
- **Percentiles:** Similar to quartiles, Percentiles separate the dataset into 100 equal parts. Percentiles are termed as P1, P2, P3,….,P100. Each part contains 1% of the data.

Continue to next slides where these concepts are illustrated with examples.

- Like median, you can sort the data in ascending or descending order, but by convention and for the ease of interpretation, we sort the data in ascending order.

Calculating Quantiles

- Formula: $\frac{i}{4} * (n + 1)th$ element in the dataset is the quartile.
- What if the formula yields an odd number?
- Then the quartile is directly not found from the dataset. It is the average of the value in the integer position of the yielded value and its next value of the position.

Example 01: Calculate the values for Q1, Q2, and Q3 from the dataset

X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Calculating Quartiles

Example 01: Calculate the values for Q1, Q2, and Q3 from the dataset

X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Solution: Given, X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Step 01: Sorted_X = [6.75, 8, 9, 10, 11, 13, 14, 14.25, 15, 18]

Step 02: For Q1, i = 1. So, $\frac{i}{4} * (n+1) = \frac{1}{4} * (10+1)$ = 2.75th data point is Q1. Since 2.75th position is not possible, so the First Quartile is the average of 2nd and 3rd data point in the sorted dataset.

Step 03: Q1 = $\frac{8+9}{2} = 8.5\ (Ans)$

Calculating Quartiles

Example 01: Calculate the values for Q1, Q2, and Q3 from the dataset

X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Solution: Given, X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Step 01: Sorted_X = [6.75, 8, 9, 10, 11, 13, 14, 14.25, 15, 18]

Step 02: For Q2, i = 2. So, $\frac{i}{4} * (n + 1) = \frac{2}{4} * (10 + 1)$ = 5.5th data point is Q2. Since 5.5th position is not possible, so the Second Quartile is the average of 5th and 6th data point in the sorted dataset.

Step 03: Q2 = $\frac{11+13}{2} = 12\ (Ans)$

Calculating Quartiles

Example 01: Calculate the values for Q1, Q2, and Q3 from the dataset

X =

Solution: Given, X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Step 01: Sorted_X = [6.75, 8, 9, 10, 11, 13, 14, 14.25, 15, 18]

Step 02: For Q3, i = 3. So, $\frac{i}{4} * (n+1)$ = $\frac{3}{4} * (10+1)$ = 8.25th data point is Q3. Since 8.25th position is not possible, so the Third Quartile is the average of 8th and 9th data point in the sorted dataset.

Step 03: Q3 = $\frac{14.25+15}{2} = 14.625\ (Ans)$

Calculating Quartiles

Example 02: Calculate the values for D1, D10, and P25 from the dataset

X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Solution: Given, X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Sorted_X = [6.75, 8, 9, 10, 11, 13, 14, 14.25, 15, 18]

Calculating D1: i = 1. $\frac{i}{10} * (n+1) = \frac{1}{10} * (10+1) = 1.1$

So, D1 is the average of 1st and 2nd data point in sorted X.

D1 = $\frac{6.75+8}{2} = 7.375\ (Ans)$

Calculating Quartiles

Example 02: Calculate the values for D1, D10, and P25 from the dataset

X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Solution: Given, X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Sorted_X = [6.75, 8, 9, 10, 11, 13, 14, 14.25, 15, 18]

Calculating D10: i = 10. $\frac{i}{10} * (n+1) = \frac{10}{10} * (10+1) = 11$

Since 11th index is not possible, D10 is the last data point in the sorted X.

D10 = 18 (Ans)

Calculating Quartiles

Example 02: Calculate the values for D1, D10, and P25 from the dataset

X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Solution: Given, X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Sorted_X = [6.75, 8, 9, 10, 11, 13, 14, 14.25, 15, 18]

Calculating P25: i = 25. $\frac{i}{100} * (n + 1) = \frac{25}{100} * (10 + 1) = 2.75$

So, P25 is the average of 2nd and 3rd data point in the sorted X.

P25 = $\frac{8+9}{2} = 8.5\ (Ans)$

Quantiles Interpretation

- Quantiles interpretation is straight forward. Lets understand from a real-life example. The scientific journals are usually ranked as Q1, Q2, and Q3 journals based on the impact factor.
- Imagine, there are 100 scientific journals on Generative AI and each of these journals have their respective impact factors.
- What do we understand by a Q1 journal then?
- It means, (if the impact factors are sorted in descending order, then) that particular journal's impact factor is within the top 25% impact factors (since Q1 divides top 25% data in the left when data is sorted in descending order).
- Similar interpretations are for percentiles and deciles too. D9 denotes that 90% of the data points smaller than D9 lies in the left of D9 (when the data is in ascending order).
- Yes, the interpretation reverses on the basis of the data is in ascending or in descending order. The choice is on the statistician regarding how he wants to interpret it.
- Note that, Q2 = D5 = P50 = Median.

Measures of Dispersion

- Measures of dispersion tell us how much the data points in a dataset differ from the mean or the central value. It also explains how much the data points are spread from each other. Higher spread or difference denotes to higher variability in the data points which has practical interpretations.
- There are 6 types of measures of dispersion. These are the topics to be discussed :

- Range
- Variance
- Standard Deviation (SD)
- Mean Absolute Deviation (MAD)
- Interquartile Range (IQR)
- Coefficient of Variation (CV)

- Variance and SD has been discussed already in the pre-requisite chapter. Also note that, while discussing IQR, we will cover outlier detection and box-plot analysis as well.

Range

- The difference between the largest and smallest data point in a univariate dataset is known as range. Assume a dataset Y = [5,1,9,3,6]. Here, the range is (9-1) = 8.
- We use range as a measure of dispersion because –
- It is easy to calculate and hence gives a quick overview on the data's overall spread and variability.
- Range is a quick measure to compare the variability between multiple dataset. The dataset having the largest range is more spread out than the other datasets.
- However, range has certain limitations.
- It is affected by outliers when the largest or smallest data point is an outlier.
- It only works with the largest and smallest values, the distribution of the data is ignored.

Mean Absolute Deviation

- The average of absolute deviations of data points in a dataset is known as the Mean Absolute Deviation (MAD). It is also described as "mean of absolute deviations".
- What is **deviations?**

- The differences of every data point from certain target values are deviations. Assume, a dataset X = [3,4,5]. Here, mean = 4.

So, deviations from the mean are: [(3 – 4), (4 – 4), (5 – 4)] = [-1, 0, 1]

- What is **absolute deviations?**

- We get absolute deviations by taking the absolute form of the deviations.

Absolute Deviations are: [1, 0, 1].

- Then, what is **Mean Absolute Deviation (MAD)?**

- Just calculate the mean of absolute deviations, and we get the Mean Absolute Deviation.

MAD = (1 + 0 + 1)/3 = 0.67

Mean Absolute Deviation

- **Example 01**: Calculate MAD for the dataset X = [6, 1.25, 7, 8.75, 4].

Solution: Given, dataset X = [6, 1.25, 7, 8.75, 4]

Step 01: Mean = $\frac{6+1.25+7+\ .75+4}{5} = 5.4$

Step 02: Deviations from mean = [(6 – 5.4), (1.25 – 5.4), (7 – 5.4), (8.75 – 5.4), (4 – 5.4)] = [0.6, -4.15, 1.6, 3.35, -1.4]

Absolute deviations = [0.6, 4.15, 1.6, 3.35, 1.4]

Step 03: Hence, Mean Absolute Deviation (MAD) = $\frac{0.6+4.15+1.6+3.35+1.4}{5} = 2.22\ (Ans)$

Mean Absolute Deviation

- **Why do we use Mean Absolute Deviation (MAD)?**

- The unit of MAD is same as the values in the dataset. If the dataset represents weights of certain material in kilograms, the unit of MAD will also be in kilograms. Hence, MAD is easier to interpret.
- MAD is more robust to outliers than Variance and Standard deviation due to the absolute values not being squared.
- Calculating MAD is easier than Standard Deviation or Variance, and hence when a quick overview on spread comparison between two datasets is necessary, MAD is the best measure of dispersion.
- MAD is a better measure for accuracy during calculating prediction errors that are continuous data in nature.
- Similarly, in quality control processes, MAD can help monitor the consistency of production by measuring the average deviation from a target value.

Interquartile Range (IQR)

- The difference between Q3 and Q1 is known as the Interquartile Range (IQR). IQR = Q3 – Q1.
- **Example 01**: Find the IQR of the dataset X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Solution: Given, X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]

Step 01: Sorted_X = [6.75, 8, 9, 10, 11, 13, 14, 14.25, 15, 18]

Step 02: For Q1, i = 1. So, $\frac{i}{4} * (n+1) = \frac{1}{4} * (10+1)$ = 2.75th data point is Q1. Since 2.75th position is not possible, so the First Quartile is the average of 2nd and 3rd data point in the sorted dataset.

Step 03: Q1 = $\frac{8+9}{2} = 8.5$

Step 04: Similarly, Q3 = $\frac{14.25+1}{2} = 14.625$

Step 05: So, IQR = Q3 – Q1 = 14.625 – 8.5 = 6.125 (Ans)

BRAC UNIVERSITY
Inspiring Excellence

Interquartile Range (IQR)

- **Why do we use Interquartile Range (IQR)?**
- IQR is less sensitive to outliers as the outliers usually do not exist between Q3 and Q1.
- IQR measures the spread of the middle 50% of the data, which is often of primary interest. Hence, its robust to skewed distributions too.
- IQR is used to detect outliers. It is the famous method of detecting outliers which was developed by John Tukey.
- IQR has a limitation.
- Just like Range, IQR does not care about the whole dataset, only prioritizes the central or more common values.

Outlier Detection

- Datasets often have values/data points that are extremely larger or smaller than the usual values. These are known as **outliers** or **extreme values**.
- John Tukey developed the formula of detecting outliers using the concepts of IQR in his book *Exploratory Data Analysis (EDA)* in 1977.
- The method is known as "IQR Method for Outlier Detection", or simply "Tukey's Fences".
- Non-outlier threshold according to the method: **(Q1 – 1.5 * IQR, Q3 + 1.5 * IQR)**. Any data point that does not fall within this threshold is considered as an outlier.
- The use of 1.5 in the method is intuitive. Because 2 would enlarge the threshold too much that even the outliers could fall within the threshold, eventually making the whole method inefficient.

Outlier Detection

- **Example 01:** Determine the low and high outlier threshold for the dataset X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75]. Does the dataset contain any outlier?

Solution: We already found, Q1 = 8.5 and Q3 = 14.625.

IQR = 6.125

Non-outlier threshold/interval = (8.5 – 1.5 * 6.125, 14.625 + 1.5 * 6.125) = (-0.68, 23.81).

So, any values less than -0.68 will be a low outlier in the dataset X, and values greater than 23.81 will be a high outlier in the dataset. (Ans)

Since, every value / data point in dataset X lies within the non-outlier interval, dataset X has no outliers. (Ans)

Box Plot

- Also known as box-and-whisker plot.
- The following **5** values are required to draw a box plot:
- Minimum Value
- Maximum Value
- Median
- Q1 (First Quartile)
- Q3 (Third Quartile)

Place the values in a graph paper to construct the box plot. (You can rename those 5 values as 3M2Q to recall them more efficiently.)

Box Plot

- **Example 01**: Construct a box plot for the dataset X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75].

Solution: Minimum value = 6.75

Maximum value = 18

Median = 12

First Quartile Q1 = 8.5

Third Quartile Q3 = 14.625

(Continue to next slide..)

Box Plot

- **Example 01**: Construct a box plot for the dataset X = [13, 8, 9, 15, 14.25, 14, 10, 11, 18, 6.75].

Solution: By placing them in a graph paper,

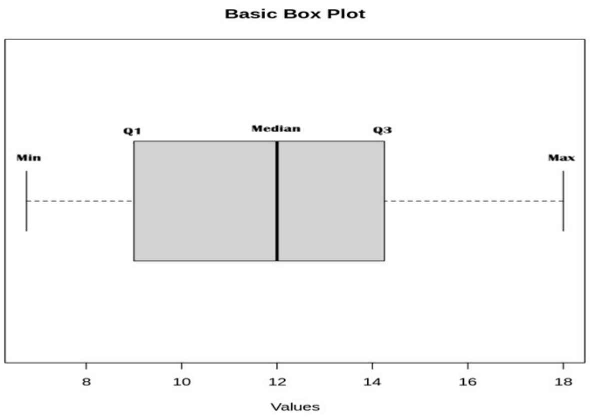


Fig 1.2: Box plot

Fig 1.2 shows the box plot that quickly visualizes the spread of data. The longer the box is horizontally, the more spread the data is. The general box plot does not contain whiskers. If you plot the low outlier threshold and high outlier threshold in the box plot, it becomes a box-and-whisker plot, which helps you to detect outliers.

Coefficient of Variation (CV)

- We get coefficient of variation by dividing standard deviation with the mean of the data and usually represent it as percentage. (If the data represents population data, while calculating standard deviation, do not use n-1 in the denominator as mentioned in pre-requisite chapter. Use n instead. For usual cases, use n-1.)
- **Example 01:** Calculate CV for the dataset Z = [2.25, 5, 1.45, 7].

Solution: Sample mean, $\bar{x} = \frac{2.25+5+\ .45+7}{4} = 3.925$

Sample SD, $s = \sqrt{\frac{(2.25-3.925)^2+(5-3.925)^2+(1.45-\ .925)^2+(7-3.925)^2}{4-1}} = 2.55$

So, CV of Z = $(\frac{S}{\bar{x}} * 100)\% = \left(\frac{2.55}{3.925} * 100\right)\% = 64.9\%\ (Ans)$

Coefficient of Variation (CV)

- **Why do we use Coefficient of Variation (CV)?**
- CV has no units and therefore is useful when comparing variability in datasets with different units of measurements. For example, if you have to compare the spread of two datasets one containing data in kilogram unit and other in meter unit, you can use CV.
- CV is practically interpretable with fluctuation rates. A higher CV denotes more spread or fluctuations in the data, which may either point to positive or negative consequences. The example is illustrated in the next slide broadly since this topic is very important to understand.

Skewness

- In the pre-requisite chapter, we learned the necessity of a normally distributed data. A data is normally distributed when it is not skewed. So, measuring skewness of a data is necessary for statistical analyses. We can measure skewness by :

- Visualizing the data by dot plot or histograms.
- Using Pearson’s First Coefficient of Skewness.
- Using Pearson’s Second Coefficient of Skewness, also known as the adjusted Fisher-Pearson standardized moment coefficient.

We will shortly discuss the 2nd and 3rd method for calculating skewness in this section.

Pearson’s First Coefficient of Skewness

- Formula: Skewness = $\frac{Mean - Mod}{Standard\ Deviation}$
- If the calculated value of skewness is positive, then the data is positively skewed (skewed to the right). Else, if the calculated value of skewness is negative, then the data is negatively skewed (skewed to the left).
- If the skewness value is 0, then the data is not skewed. The data is normally distributed/ symmetrically distributed.
- If the value is between -0.5 to 0.5, then the data is considered to be approximately normally distributed.
- Values between -2 and -1 or 1 and 2 suggests moderate skewness.
- Values between -1 and -0.5 or 0.5 and 1 suggests slight skewness.
- Values less than -2 or greater than 2 indicates high skewness.

Pearson's First Coefficient of Skewness

- **Example 01**: Assume a dataset X = [1, 5.5, 5.5, 3, 3.25, 7]. Calculate first coefficient of skewness and comment the distribution.

Solution: Skewness = $\frac{Mean - Mode}{Standard\ Deviation}$

Mean = 4.20

Mode = 5.5

SD = 2.18

So, Skewness = $\frac{4.20-5.5}{2.18} = -0.59$ (Ans)

Since the skewness is in between -1 to -0.5, the data is **slightly skewed to the left**. (Ans)

Adjusted Fisher-Pearson Method

- The method is mostly used to calculate skewness because it is more useful.
- Calculating median is easier than calculating mode. Strong programming languages like R and Python also do not have a direct function to calculate mode.
- Formula: Skewness = $\frac{3(Mean-Median)}{Standard\ Deviation}$
- The thresholds for interpreting skewness is the same as First Coefficient of Skewness method.

Adjusted Fisher-Pearson Method

- **Example 01**: Assume a dataset X = [2.75, 5.5, 5.5, 5, 5.25, 9]. Calculate first coefficient of skewness and comment the distribution.

Solution: Skewness = $\frac{3(Mean-Median)}{Standard\ Deviation}$

Mean = 5.5

Median = 5.375

SD = 2.006

So, Skewness = $\frac{3(5.5-5.375)}{2.006} = 0.18$ (Ans)

Since the skewness is very close to 0 (between -0.5 to 0.5), the data is **approximately normally distributed**. The positive value denotes that the data is **very slightly skewed to the right**.

Stem-and-leaf Plot

- Stem-and-leaf plot is the one of the most useful and easy method for:
- Understanding the underlying distribution of the data,
- Detecting outliers
- Sorting the data
- Estimating median and mode
- The steps to make a stem-and-leaf plot is described in the next slide.

Stem-and-leaf Plot

- **Example 01**: Create a stem-and-leaf plot of the data X = [60, 65, 70, 55, 52, 55, 55, 52, 60, 65, 67, 100, 110, 105, 44, 45, 44, 55, 57, 109]

Solution

Step 01: Write down the first digits of data points in such way that there is only one digit left in the right of the data point as shown in the following. These first digits are known as **stems.**

4

5

6

7

8

9

10

Stem-and-leaf Plot

- **Example 01**: Create a stem-and-leaf plot of the data X = [44, 45, 44, 60, 65, 70, 55, 52, 55, 55, 52, 60, 65, 67, 100, 110, 105, 55, 57, 109]. Estimate median, mode, and comment on distribution.

Solution

Step 02: Place the last digit of every data point to where it belongs. These last digits are **leaves**.

4 : 4, 5, 4

5 : 5, 2, 5, 5, 2, 5, 7

6 : 0, 5, 0, 5, 7

7 : 0,

8 :

9 :

10 : 0, 5, 9

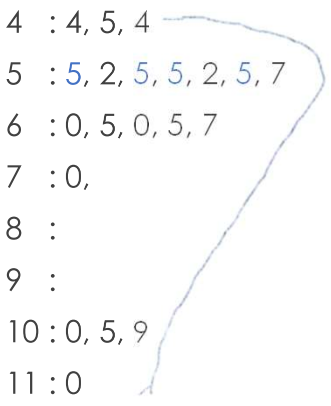
11 : 0

Stem-and-leaf Plot

- **Example 01**: Create a stem-and-leaf plot of the data X = [44, 45, 44, 60, 65, 70, 55, 52, 55, 55, 52, 60, 65, 67, 100, 110, 105, 55, 57, 109]. Estimate median, mode, and comment on distribution.

Solution

Step 02: Place the last digit of every data point to where it belongs. These last digits are **leaves**.



4 : 4, 5, 4
5 : 5, 2, 5, 5, 2, 5, 7
6 : 0, 5, 0, 5, 7
7 : 0,
8 :
9 :
10 : 0, 5, 9
11 : 0

As we can see, most data points are scattered around in the left of the number line, so in the **right** there are a few data points. Hence, the distribution is rightly skewed. (Ans)

Under the stem 5, the number 5 appeared the most. So, the **mode** is 55. (Ans)

The size of X is 20. So, the median will be the average of 10th and 11th datapoint in the sorted dataset. From the plot, we get,

Sorted X = [44, 44, 45, 52, 52, 55, 55, 55, 55, 57,60, 60, 65, 65, 67, 70, 100, 105, 109, 110]

So, **median** = (57+60)/2 = 58.5 (Ans)

Appendices

- Calculating central tendencies and measures of dispersion for grouped data:

https://drive.google.com/file/d/13KJM_5MVsT1sxZQJgyliu_RviZywFIHG/view?usp=sharing

https://drive.google.com/file/d/1wcUEml6vgS5p8T6L4SDAtsx_hFjlSgiZ/view?usp=drive_link

Appendices

- Drawing a histogram: How To Make a Histogram Using a Frequency Distribution Table – YouTube
- Drawing a frequency polygon: How To Make a Frequency Polygon (youtube.com)
- Drawing a ogive curve: Statistics Grade 11: Ogive curve (youtube.com)
- Bar chart: What is a Bar Chart? (youtube.com)
- Creating cumulative relative frequency distribution: How To Make a Cumulative Relative Frequency Table (youtube.com)

STA201: Elements of Statistics and Probability
Department of MNS, BRAC University

Chapter 02: Probability

KAZI SAKIB HASAN (24341237)

COMPUTER SCIENCE PROGRAM, BRAC UNIVERSITY

kazi.sakib.hasan@g.bracu.ac.bd

Brief Contents

- PROBABILITY TERMINOLOGIES
- CALCULATING PROBABILITY
- BINOMIAL DISTRIBUTION
- HYPERGEOMETRIC DISTRIBUTION
- GEOMETRIC DISTRIBUTION
- POISSON DISTRIBUTION
- MARGINAL PROBABILITY, CONDITIONAL PROBABILITY, ADDITIONAL RULE
- FINITE SAMPLE TRANSFORMATION
- MUTUALLY EXCLUSIVE EVENTS
- INDEPENDENT AND DEPENDENT EVENTS
- CALCULATING OUTCOMES

Probability

- The definition of probability lies in the term itself. Probability is the possibility or the likelihood of an event's occurrence which is denoted by a number ranging from 0 to 1, where lower magnitude defines low possibility and higher magnitude defines higher possibility. So, $0 \leq P(E) \leq 1$.
- Probability can be represented as **percentage.**
- The best way to understand probability concepts is by solving mathematical problems by creating a general method that can help us to solve any type of probability problems. The general method we can use do so is **Finite Sample Transformation**.
- This chapter is highly focused on calculating probabilities using the method mentioned above, instead of using traditional theoretical methods.

Terminologies

- You need to know about some terminologies that are studied in probability. Consider the following scenario.

  A dice is rolled on the board and the player got a 6.

- **Experiment**: Rolling the dice is an experiment.
- **Outcome**: Getting 6 is the outcome. Also known as simple event or event.
- **Sample Space**: Set of every possible outcomes is the sample space. Sample Space, S = {1, 2, 3, 4, 5, 6}. The elements of sample space are **sample points.**
- **Event**: An event is a subset of the sample space. For example, {2,4,6} is an event.
- **Simple Event**: When the Event subset contains only one element, it is a simple event. Getting 6 is a simple event, getting 1 is a simple event etc.
- **Compound Event**: Getting multiple events as outcome is a compound event.

Calculating Probability

- Eight methods of calculating probability based on different scenarios:
- Classical Probability Approach
- Relative Frequency Approach
- Subjective Probability Approach
- Binomial Distribution
- Hypergeometric Distribution
- Geometric Distribution
- Poisson Distribution
- Finite Sample Transformation

Classical Probability

- **When to use classical probability approach?**

- When the length of sample space is known.
- The length of desired outcome's event subset is known.

Formula: $Classical\ probability, P(A) = \frac{\text{Length of event } subset}{Length\ of\ sample\ space\ set}$

Example 01: A fair coin is tossed. Find the probability of getting a tail.

Solution : Checklist:

✓ The sample is known. S = {Head, Tail}. Length of S = 2

✓ Desired outcome is getting a Tail. So, event, E = {Tail}. Length of E = 1

Therefore, we can apply the classical approach formula to count the probability.

So, P(getting a tail) = $\frac{1}{2} = 0.5\ (Ans)$

- Note that, a synonym of length is "size". If S = {A,B,C,F,}, the length, or the size of S is 4 because there are 4 elements in S.

Classical Probability

- **Example 02**: What is the probability of getting an odd number in one roll of a dice?

Solution: Checklist:

✓ Size of Sample Space is known. S = {1,2,3,4,5,6}. So, size of S = 6.

✓ Desired outcome is an odd number. So, E = {1,3,5}. So, size of E = 3

Therefore, we can apply the classical approach formula to count the probability.

P(getting an odd number) = $\frac{3}{6} = 0.5\ (Ans)$

BRAC UNIVERSITY

Inspiring Excellence

Classical Probability

- **Example 03**: There are 20 balls in a bag, among them 12 are red. If a randomly ball is selected, what is the probability of the ball is red?

Solution: Checklist:

✓ Size of Sample Space is known. S = $\{Normal\ Ball^1, Normall\ Ball^2, \ldots, Normal\ Ball^8, Red\ Ball^1, Red\ Ball^2, \ldots Red\ Ball^{12}\}$. So, sample space size = 20.

✓ Desired outcome is a red ball. So, event = $\{Red\ Ball^1, Red\ Ball^2, \ldots Red\ Ball^{12}\}$. So, size of E = 12.

P (getting a red ball) = $\frac{12}{20} = 0.6\ (Ans)$

Classical Probability

- **Example 04:** In a group of 400 boys, 130 had a relationship once in their life. If a boy is randomly selected, what is the probability that he had a relationship once?

Solution: Checklist:

✓ Length of Sample Space is known. S = $\{NormalBoy^1, NormalBoy^2, .. NormalBoy^{270}, RelBoy^1, RelBoy^2, .. RelBoy^{130}\}$. Length of S = 400.

✓ Desired outcome is the boy had a relationship once. So, E = $\{RelBoy^1, RelBoy^2, .. RelBoy^{130}\}$. Length of E = 130

So, P(boy had a relationship) = $\frac{130}{400} = 0.325\ (Ans)$

BRAC UNIVERSITY

Inspiring Excellence

Classical Probability

- **Example 04:** In a group of 400 boys, 130 had a relationship once in their life. If a boy is randomly selected, what is the probability that he had a relationship once?

Alt Solution: Probabilities are nothing but percentages. The percentages of boys had a relationship = probability of getting a boy who had a relationship.

So, $\frac{130}{400} * 100 = 32.5\% \ (Ans)$

Or, $\frac{130}{400} = 0.325 \ (Ans)$

Note: This alternative method is applicable for every classical probability problem shown before. We will use this alternative method in the future sections.

Classical Probability

- **Example 05:** In a group of 400 boys, 130 had a relationship once in their life. If a boy is randomly selected, what is the probability that he **never** had a relationship once?

Solution: Among 400 boys, 130 had a relationship once. So, (400-130) = 270 boys **never** had a relationship once.

So, P(boy never had a relationship) = 270/400 = 0.675 (Ans)

Alt Solution: From example 04, P(boy had a relationship) = 0.325

So, P(boy never had a relationship) = 1 - 0.325 = 0.675 (Ans)

Note: If A is an event, then $\bar{A}$ is the **complementary event** of A. It includes all the outcomes for an experiment that are not in A. P($\bar{A}$) = 1 – P(A)

Relative Frequency

- **When to use relative frequency approach?**

- The length of sample space is unknown.
- Therefore, the event subset is also unclear.

Example 01: What is the probability that a randomly selected student will get A in STA201 course under SBBH?

- Solving this problem is not possible. Because the sample space is unknown. So, we need to conduct a experiment first. We can randomly sample his previous students' earned grade records to come to a conclusion.

Relative Frequency

- **Example 01:** What is the probability that a randomly selected student will get A in STA201 course under SBBH? Assume that, a sample of his previous 1000 students have been selected, among them 450 got A.

Solution: Applying the formula of classical probability approach, we get,

P (getting A under SBBH) = $\frac{450}{1000} = 0.45\ (Ans)$

- **Example 02:** What is the probability that the next car coming out from "Palki Motors" is White? A sample of 500 cars is already taken and it is found that 430 cars were White.

Solution: P (getting a White car) = $\frac{430}{500} = 0.86\ (Ans)$

Relative Frequency

- Relative Frequency approach is on the basis of Law of Large Numbers. You need to conduct an experiment for a long time and record the outcomes to get the relative frequency of that outcome. Only then, the probability will be very closer to the real value of probability/ theoretical probability.
- Carl Pearson once tossed a fair coin 24000 times and got head 12012 times. The relative frequency of getting heads in a fair coin is thus $\frac{12012}{24000} = 0.5005$, which is almost the same as the theoretical probability 0.5.
- Now, how long is long enough?
- An exact answer to this question is not possible. It varies according to the type of experiment. But typically, 50 trials to 100 trials can be a good choice if time restriction is concerning.
- But also note that, the more trial you conduct, the more accurate the relative frequency is to the theoretical probability.

Relative Frequency

- **Example 03**: You want to determine the probability of Far Cry 4 crashes while you play the game without disconnecting Wi-Fi. How can you do so?

Solution: Connect your Wi-Fi and run the game for a couple of hours. Imagine, you ran the game 50 times each time for 2 hours and found that your game crashed 25 times.

So, P (FC4 crashes with Wi-Fi) $= \frac{25}{50} = 0.5$

What decision should you take?

Since there is 50-50 chance of your game being crashed while playing with Wi-Fi, so you should disconnect the Wi-Fi before playing the game for a better gaming experience.

Subjective Probability

- A belief that classical and relative frequency approach do not determine the correct value of probability. Consider the example of a random student getting A under SBBH.
- If there were only brilliant students in that sample, then the relative frequency of getting A would increase significantly. Conversely, sample with less brilliant students would decrease the relative frequency.
- Therefore, sometimes its not easy to determine the probability of an event absolutely correctly. In such cases, we infer the probability.
- This inference on probability value varies person to person, and hence the probability is known as subjective probability.

Subjective Probability

- **Example 01:** Syn and Ash are in relationship with each other. How can they determine the probability of whether they can marry each other in the future?

Solution: There are multiple approaches to calculate the probability.

- Maybe they follow a certain pattern in their relationship. They can list couple who followed the same and could marry each other. Thus, they will get a relative frequency of marrying each other.
- Maybe they are also teenage. So, they can list teenage couples who successfully married each other. The relative frequency will answer their question.
- Maybe they are also from different religions. So, they again can list couples with different religions and calculate the relative frequency of marriage.

Thus, in each cases, there will be a different probability of marrying each other in the future. This phenomenon is known as the subjective probability.

Binomial Distribution

- Binomial probability distribution is a discrete probability distribution. So do the geometric, and Poisson distribution. Discrete probability distribution will be discussed later.
- **When to use binomial distribution?**

- In a random experiment where there can be only two outcomes: **success** and **failure**.
- Two outcomes have their own **constant** probability of occurring.
- If you want to find out the probability of **k** successes in **n** trials, you can use binomial distribution.

Formula: $P(x = k) = \binom{n}{k} p^k . (1 - p)^{n-k}$

Here, n = Number of trials.

k = Desired number of success.

p = Probability of success.

$1 - p$ = Probability of failure.

Binomial Distribution

- **Calculating expected value (mean) and SD for Binomial Distribution**

1. Calculating expected value of binomial distribution: **Expected value,** $\mu = \boldsymbol{np}$**.** It tells us the expected number of **successes** we will get in $\boldsymbol{n}$ trials. Here, $\boldsymbol{p}$ is the constant probability of success.
2. Calculating SD of binomial distribution: **Standard deviation, σ =** $\sqrt{\boldsymbol{np(1-p)}}$ It shows us the interval into which the expected successes can fall: **(**$\mu \pm$ **σ)**

Binomial Distribution

- **What is the meaning of constant probability of success?**

- Imagine a box with 16 red pencils and 4 blue pencils. Assume, your eyes are closed and if you randomly pick a blue pencil, then it is a **success.** So, currently the success probability is 4/20 = 0.2
- If you pick up a blue pencil in your 1st trial **without replacement**, now there are 3 blue pencils in the box and 16 red pencils. Now, the success probability becomes 3/19 = 0.15. Before it was 0.2.

So, in this scenario, the probability of success is not constant. It depends on the trials. For each trial, the probability changes. We cannot use binomial formula in such cases. In such cases, we use the **hypergeometric distribution.**

Binomial Distribution

- **What is the meaning of constant probability of success?**

- But what happens if you had picked up the blue pencil in your 1st trial **with replacement**? Then, even after your 1st trial, the probability of success would remain constant. Because there would be 4 blue pencils and 16 red pencils in that box for your second trial. So, the probability of success would be still 0.2 in your 2nd trial. We can use binomial formula in such cases.
- So, if the probability of success is the same for every trial, then we can use binomial formula to calculate probability. If the sampling is done **with replacement**, feel free to use the binomial formula. Else, if the sampling is **without replacement**, do not use the formula.

Binomial Distribution

- **What is the meaning of constant probability of success?**

- Furthermore, there are cases where the sampling is not done **with replacement**, yet the probability of success remains constant in every trial. For example :

1. Infinite population cases like manufacturing industry. No matter how many products you choice as trials **without replacement**, the probability of success will always remain the same since there are literally infinite products in that industry. Picking up several products would not wrangle this value. The probability of success is nothing but the **relative frequency** of successful products.
2. Cases where sampling is not done. A footballer's probability of successfully scoring goals in free kicks is calculated by the **relative frequency** of his previous matches. As a result, this relative frequency (probability of success) always remains constant.

In short, if the probability of success is a relative frequency of long term experiment, then the probability is always constant in each trial.

Binomial Distribution

- **Sampling with replacement:** Assume that you need to sample 5 students from 30 students in your classroom randomly (closing your eyes).

- You pick the 1st one. You recorded his ID number (X01), and said him to get back to the classroom again.
- When you are about to select the second student, the 1st one (X01) can be selected again, since you are randomly picking students.

This type of sampling is known as sampling with replacement, where you first draw an sample element, then throw it back to the population, and then again draw another sample element, again throw back the second element to the population, and so on so forth.

Binomial Distribution

- **Sampling without replacement:** Assume that you need to sample 5 students from 30 students in your classroom randomly (closing your eyes).

- You pick the 1st one. You recorded his ID number (X01), and said him to leave the classroom.
- When you are about to select the second student, the 1st one (X01) cannot be selected again, because he is already out of the room.

This type of sampling is known as sampling without replacement, where you first draw an sample element, then do not throw it back to the population, and then again draw another sample element, again do not throw back the second element to the population, and so on so forth.

Binomial Distribution

- **Example 01:** If a fair coin is tossed 20 times, what is the probability of getting 15 heads?

Solution: Checklist:

✓ Experiment with only two outcomes: S = {Head, Tail}

✓ We want to find the probability of getting heads. So, success = getting head. Failure = getting tail.

✓ The probability of getting head and tail is 0.5, and this value is constant.

So, we can apply the binomial distribution formula to calculate the probability.

Given, n = 20, k = 15

P(x=15) = $\binom{20}{15}0.5^{15}.(1-0.5)^{20-15} = 0.014\ (Ans)$

Binomial Distribution

- **Example 02:** 5% resistors in an electrical engineering lab is defected. If 3 resistors is selected, what is the probability that exactly 2 wire is defected?

Solution: Checklist:

- ✓ Experiment with only two outcomes. S = {Defected, Alright}
- ✓ We want to find out the probability of defection. So, success = the resistor is defective. Failure = the resistor is alright.
- ✓ The probability of success (p) is 0.05, and the probability of failure (1-p) is (1-0.05). These values are constant.

Given, n = 3, k = 2

So, applying binomial formula, we get,

P(x=2) = $\binom{3}{2}0.05^2.(1-0.05)^{3-2} = 0.007125\ (Ans)$

Binomial Distribution

- **Example 03:** 5% resistors in an electrical engineering lab is defective. If 3 resistors is selected, what is the probability that **at most** 2 wire is defective?

Solution: Checklist:

- ✓ Experiment with only two outcomes. S = {Defective, Alright}
- ✓ We want to find out the probability of defection. So, success = the resistor is defective. Failure = the resistor is alright.
- ✓ The probability of success (p) is 0.05, and the probability of failure (1-p) is (1-0.05). These values are constant.

Given, n = 3, maximum k = 2

So, applying binomial formula, we get,

P(x ≤ 2) = P(x=0) + P(x=1) + P(x=2)

$= \binom{3}{0}0.05^0.(1-0.05)^{3-0} + \binom{3}{1}0.05^1.(1-0.05)^{3-1} + \binom{3}{2}0.05^2.(1-0.05)^{3-2} = 0.85 + 0.135 + 0.007125 = 0.99\ (Ans)$

Binomial Distribution

- **Example 04:** Assume that a classifier machine learning model classifies samples with 85% accuracy. Suppose, 7 samples are taken for classification using the model. What is the probability of –

(a) Getting 5 accurate classification?

(b) Getting **at most** 3 accurate classification?

(c) Getting **at least** 4 accurate classification?

Binomial Distribution

- **Solution A:** Given, n = 7 and k = 5

  p = 0.85

  So, P(x=5) = $\binom{7}{5} * 0.85^5 * (1 - 0.85)^{7-5} = 0.20\ (Ans)$

- **Solution B:** P(x ≤ 3) = P(x=0) + P(x=1) + P(x=2) + P(x=3) = 0+0+0.001+0.01 = 0.011 (Ans)

- **Solution C:** P(x ≥ 4) = P(x=4) + P(x=5) + P(x=6) + P(x=7) = 1 – P(x ≤ 3)

  = 1-0.011

  = 0.989 (Ans)

- Note, P(x=0) + P(x=1) + P(x=2) + …. + P(x=k) = 1. This statement is true for geometric distribution and Poisson distribution as well.

BRAC UNIVERSITY

Inspiring Excellence

Binomial Distribution

- **Example 05:** Synthia wants to find out the probability of a footballer failing to goal 3 to 5 times in 5 shots of penalty. How can she do so?

Solution: First of all, she needs to define success and failure in the experiment. Here, she wants to find out the probability of missing goals in penalty. So, **success** = failing to score a goal, and **failure** = success to score a goal.

Secondly, she needs to calculate the relative frequency of penalty failure for that player. For this, she can rewatch that player's previous penalty shoot outs and can calculate the relative frequency of penalty failure. This relative frequency of failure is the constant probability of success in this experiment.

Lastly, she can apply the binomial formula to calculate the probability.

Binomial Distribution

- **Example 05:** Syn finds out that the player shot 45 penalties in his career and among those he missed 25 shots. Calculate the probability that the player will miss 3 to 5 shots in 5 trials. If a win is must required in that match, does Syn should deploy that player in tie-breaker round?

Solution: Constant probability of success, p = 25/45 = 0.56

P(3≤ x ≤ 5) = P(x=3) + P(x=4) + P(x=5)

= $\binom{5}{3}0.56^3.(1-0.56)^{5-3} + \binom{5}{4}0.56^4.(1-0.56)^{5-4} + \binom{5}{5}0.56^5.(1-0.56)^{5-5}$

= 0.339 + 0.216 + 0.05 = 0.605 (Ans)

There is 60.5% chance that the player will miss 3 to 5 shots in 5 trials. So, its risky to deploy that player in tie-breaker round. (Ans)

Binomial Distribution

- **Example 06:** An office building has two fire detectors. The probability is 0.02 that any fire detector of this type will fail to go off during a fire. Find the probability that both of these fire detectors will fail to go off in case of a fire.

Solution: Here, **success** = fire detector will fail to go off.

Success probability, p = 0.02

Number of total trials, n = 2

Desired number of success, k = 2

So, P(x=2) = $\binom{2}{2}0.02^2.(1-0.02)^{2-2} = 0.0004\ (Ans)$

Binomial Distribution

- **Additional Example :** I once randomly sampled 200 girls from BRAC University that belongs to a certain population and sent them friend requests. 55 of them accepted the request within the follow up period. Now, answer the following questions.

A. Calculate the probability of acceptance and assume that the probability is constant. Now, calculate what is the probability that I will be accepted by 4 girls within two weeks if I send friend requests to 10 girls?

B. What is the probability that 4 to 6 girls will accept the friend request?

C. If I send friend request to 150 girls, how many girls do you think will accept? (Calculate expected value)

D. The approximation will vary. Determine the interval of these variations and interpret (calculate standard deviation).

E. Is it unusual that from the sample of 150 girls, 60 girls will accept my friend request?

F. If I send friend request to 100 girls, what's the probability that at most 50 of them will accept?

Binomial Distribution

In Additional Example, the questions A and B is related to calculating probability.

Question C, D and is related to Discrete Random Variables that we will study in later chapters.

Question E and F is related to the topic Normal Distribution that we will study in later chapters.

However, I still solved every question so that I can cover the entire Binomial Distribution chapter here. But your focus should be on solving problems like question A and question B right now.

Binomial Distribution

- **Solution A:** Here, success = Friend request being accepted.

Failure = Friend request being rejected.

Given, n = 10 and k = 4

Probability of success, p = 55/200 = 0.275 (Ans)

Now, P(x=4) = $\binom{10}{4}0.275^4.(1-0.275)^{10-4}$ = 0.17 (Ans)

- **Solution B:** P(4 ≤ x ≤ 6) = P(x=4) + P(x=5) + P(x=6)

= $\binom{10}{4}0.275^4.(1-0.275)^{10-4}+\binom{10}{5}0.275^5.(1-0.275)^{10-5}+\binom{10}{6}0.275^6.(1-0.275)^{10-6}=0.17+0.07+0.02=0.26\ (Ans)$

BRAC UNIVERSITY

Inspiring Excellence

Binomial Distribution

- **Solution C:** Expected value, μ = np = 150 * 0.275 = 41.25

Approximately 41 girls will accept the request. (Ans)

Alternative Solution C: We can use the unitary method.

From 200 girls, 55 accepted.

From 1 girl, 55/200 will accept.

From 150 girl, (55/200)*150 = 41.25 girl will accept.

So, approximately 41 girl will accept the friend request. (Ans)

Binomial Distribution

- **Solution D:** Standard deviation, σ = $\sqrt{np(1-p)} = \sqrt{150 * 0.275(1-0.275)} = 5.46$

Interval of these variations = (41-5.46, 41+5.46) = (35.53, 46.46)

41 girls are expected to accept the friend request with a standard deviation of 5.46 (~5). It means, in some cases 5 girls can vary to accept or reject the request. (Ans)

- **Solution E:** We expected that 41 girls would accept the request. But we are observing that 60 girls accepted the request.

z-score of observation = $\frac{observed-expect}{SD} = \frac{60-41}{5.46} = 3.46 > 2$

From the empirical rule we know that 95% of the data falls between 2 SD from the mean. The rest of the data can be considered unusual. Therefore, if the absolute z-score of the observation surpasses 2, we can say that the observation is unusual. So, it is unusual that 60 girls would accept my friend request. (Ans)

BRAC UNIVERSITY

Inspiring Excellence

Binomial Distribution

- **Solution F:** P(x ≤ 50) = P(x-0) + P(x=1) + P(x=2) + P(x=3) + ….. + P(x=50).

The pen and paper process is extremely rigorous as you can see. So, we can use two special methods to solve the problem.

1. Using statistical software like R and Python.
2. Using normal distribution table.

Since, the usage of software is not allowed in STA201, and lab is also not included in the course, we cannot use the first method. But we can use the second method. We will learn it when we discuss about Normal Distribution. For now, I am showing you both methods, in case you are curious.

Binomial Distribution

- **Solution F:** Expected value, μ = np = 100 * 0.275 = 27.5

Standard deviation, σ = $\sqrt{np(1-p)} = \sqrt{100 * 0.275(1-0.275)} = 4.46$

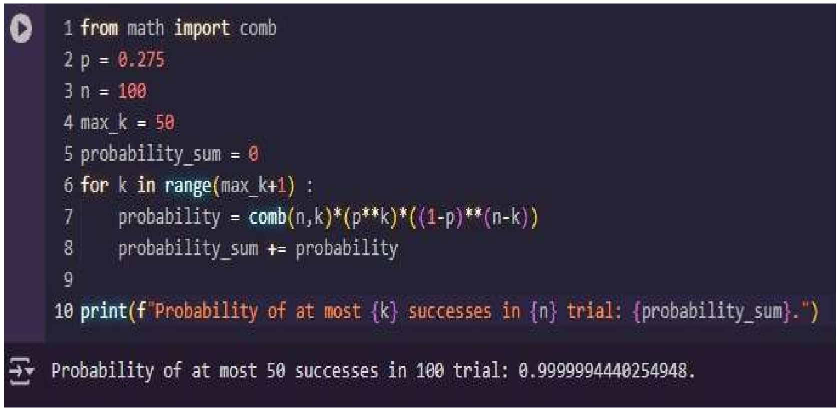
$z_{50} = \frac{50-27.5}{4.46} = 5.04$

$P(z \leq 5.04) = 0.99$ (Ans)

Note: You would not understand this approach without studying normal distribution.

Binomial Distribution

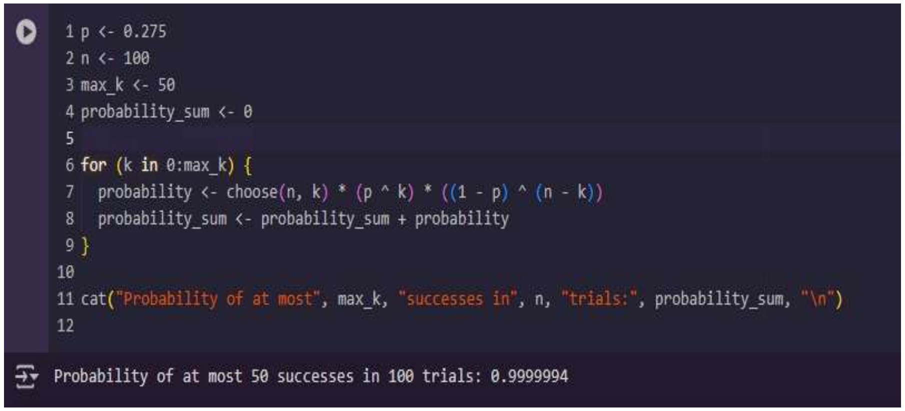
Solution F: Using statistical software (Python and R),



```python
from math import comb
p = 0.275
n = 100
max_k = 50
probability_sum = 0
for k in range(max_k+1) :
    probability = comb(n,k)*(p**k)*((1-p)**(n-k))
    probability_sum += probability

print(f"Probability of at most {k} successes in {n} trial: {probability_sum}.")
```

```
Probability of at most 50 successes in 100 trial: 0.9999994440254948.
```

Code Snippet 2.1: Python approach (procedural)



```r
p <- 0.275
n <- 100
max_k <- 50
probability_sum <- 0

for (k in 0:max_k) {
  probability <- choose(n, k) * (p ^ k) * ((1 - p) ^ (n - k))
  probability_sum <- probability_sum + probability
}

cat("Probability of at most", max_k, "successes in", n, "trials:", probability_sum, "\n")

```

```
Probability of at most 50 successes in 100 trials: 0.9999994
```

Code Snippet 2.2: R approach (procedural)

Hypergeometric Distribution

- **When to use Hypergeometric Distribution?**

- To find the probability of **k** successes in **n** trials.
- The probability of success in each trial is not constant.

Formula: $P(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$

Here, N = total number of elements in the population.

K = Number of successes in the population.

N – K = Number of failures in the population.

n = Number of trials (sample size).

k = Number of successes in n trials.

n – k = Number of failures in n trials.

Hypergeometric Distribution

- **Calculating expected value and SD for Hypergeometric Distribution**

Expected value, $\mu = \frac{nK}{N}$. This represents the average number of success items we expect to find in our sample of size $\boldsymbol{n}$ drawn from the population.

Standard deviation, σ = $\sqrt{\frac{nK}{N}(1-\frac{K}{N})(\frac{N-n}{N-1})}$ It shows us the interval into which the expected successes can fall: $(\mu \pm \sigma)$.

Hypergeometric Distribution

- **Example 01:** A box contains 20 DVDs, 4 of which are defected. If two DVDs are selected at random, what is the probability that both are defective?

Solution : Here, **success** = the item is defective.

Given, N = total number of DVDs = 20

K = Number of defective DVDs in the population = 4

n = Number of DVDs drawn as sample (trial) = 2

k = Number of defective DVDs in n trials = 2

So, $P(2) = \frac{\binom{4}{2}\binom{20-2}{2-2}}{\binom{20}{2}} = 0.0316\ (Ans)$

Geometric Distribution

- **When to use geometric distribution?**

- To count the probability of getting first success in nth trial.
- Probability of success and failure is constant.

Formula: $P(n) = (1-p)^{n-1}.p$

This concept will be relatively easier to you to understand, since you already studied Binomial Distribution.

Geometric Distribution

- **Calculating expected value and SD for Geometric Distribution**

Expected value, $\mu = \frac{\mathbf{1}}{\boldsymbol{p}}$. It represents the expected number of trials to be performed to get the first success.

SD, $\boldsymbol{\sigma}$ **=** $\sqrt{\frac{\mathbf{1-p}}{\boldsymbol{p}^{2}}}$. It shows us the interval into which the expected successes can fall: ($\mu \pm \sigma$).

Geometric Distribution

- **Example 01**: What is the probability of getting a tail in 7th coin toss?

Solution: Checklist:

- ✓ The experiment have two outcomes. S = {Head, Tail}
- ✓ The goal of this experiment is to calculate probability of getting first success in nth trial. Here, **success** = getting tail. **Failure** = getting head.
- ✓ Probability of getting tail in a coin toss 0.5, and this value is constant.

Given, n = 7.

So, P(7) = $(1 - 0.5)^{7-1}.0.5 = 0.0078 \text{ (Ans)}$

Geometric Distribution

- **Example 02**: Let us consider the previous scenario from Binomial Distribution. The probability (assuming its constant) of my friend request getting accepted by girls is 0.275. Now, answer the following questions.

A. What is the probability of getting accepted at 7th request?

B. What is the probability of getting accepted in the first 3 requests?

C. How many friend requests should I expect to send to get the first acceptance (calculate expected value)?

D. Calculate variance and standard deviation. Interpret standard deviation in this context.

Geometric Distribution

- **Solution A:** Here, **success** = getting accepted. **Failure** = getting rejected.

Given, n = 7 and p = 0.275

So, P(7) = $(1-0.275)^{7-1} * 0.275 = 0.03\ (Ans)$

- **Solution B:** P(n ≤ 3) = P(n=1) + P(n=2) + P(n=3)

= $(1-0.275)^{1-1} * 0.275 + (1-0.275)^{2-1} * 0.275 + (1-0.275)^{3-1} * 0.275$

= 0.275 + 0.19 + 0.14 = 0.605 (Ans)

- **Solution C:** E(x) = 1/p = 1/0.275 = 3.63

So, It is expected that I should send approximately 4 friend requests to get the first acceptance.

Geometric Distribution

- Alternative Solution C: We can use the unitary method.

(0.275*100) = 27.5

27.5 friend requests are accepted if 100 requests are sent

1 friend request is accepted if 100/27.5 requests are sent

= 3.63 requests ~ 4 requests (Ans)

So, It is expected that I should send approximately 4 friend requests to get the first acceptance. (Ans)

Geometric Distribution

- **Solution D:** Var(x) = $\frac{1-p}{p^2} = \frac{1-0.275}{(0.275)^2} = 9.58$ (Ans)

SD(x) = $\sqrt{9.58} = 3.09$ (Ans)

From C, expected value was 3.63.

So, interval for first success = (3.63 – 3.09, 3.63+3.09) = (0.54, 6.72) ~ (1, 7)

Approximately 4 trials are required to get the first success with an SD of 3.09, which means the first success is predicted to be achieved within 1 to 7 trials. (Ans)

BRAC UNIVERSITY
Inspiring Excellence

Poisson Distribution

- **When to use Poisson Distribution?**

- When a certain event occurred (occurrences) several times within a time interval, and you want to find the probability of occurrences in the next time interval.

Formula: $P(x) = \frac{e^{-\lambda}.\lambda^{x}}{x!}$

Here, λ = Average number of occurrences in an interval

e = Euler's number whose value is 2.718

Poisson Distribution

- **Calculating expected value and SD for Poisson Distribution**

Expected value, $\mu = \lambda$. This is the average number of occurrences or events that are expected to happen in the given interval

SD, $\sigma = \sqrt{\lambda}$. It shows us the interval into which the expected values can fall: $(\mu \pm \sigma)$.

Poisson Distribution

- **Example 01**: A certain Instagram influencer receives 9 message requests in a hour on average. Find the probability that she receives exactly 13 messages in a hour.

Solution: Given that, λ = 9 and x = 13

P(x=53) = $\frac{e^{-9}.9^{13}}{13!}$ = 0.05 (Ans)

- **Example 02**: What's the probability that she receives exactly 7 messages in 3 hours?

Solution: For 1 hour, the value for λ = 9

For 3 hours, the value for λ = 9*3 = 27

P(x = 48) = $\frac{e^{-27}.27^{7}}{7!} = 0.0000003\ (Ans)$

Poisson Distribution

- **Example 03**: A certain Instagram influencer receives 5 message requests in a hour on average. Find the probability that he receives at most 3 messages in a hour.

Solution: P(x ≤ 3) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)

$$= \frac{e^{-5}.5^0}{0!} + \frac{e^{-5}.5^1}{1!} + \frac{e^{-5}.5^2}{2!} + \frac{e^{-5}.5^3}{3!} = 0.006 + 0.03 + 0.08 + 0.14 = 0.256 \text{ (Ans)}$$

- **Example 04:** Find the probability that he receives at least 4 messages in a hour.

Solution: P(x ≥ 4) = 1 – P(x ≤ 3) = 1 – 0.256 = 0.744 (Ans)

Key Differences

- Beginner students in statistics often get confused on when to use binomial distribution, geometric distribution, and Poisson distribution. Remember that –
- Use **binomial distribution** when you are trying to find the probability of **k successes in n trials.**

- Formula: $P(x = k) \ = \binom{n}{k} p^k . (1 - p)^{n-k}$

- Use **geometric distribution** when you are trying to find the probability of **first success in nth trial.**

- Formula: $P(n) = (1 - p)^{n-1} . p$

- Use **Poisson distribution** when the average number of occurrences is given for an time interval, and you want to find the probability of that event's occurrence a given number of times.

- Formula: $P(x) \ = \frac{e^{-\lambda} . \lambda^x}{x!}$

Marginal Probability

- Before learning about Finite Sample Transformation, we need to learn about marginal probability and conditional probability. Calculating these are easy, because you can use only **classical approach** to solve these problems.
- In this section, we will learn about marginal probability. In the next section, we will discuss conditional probability briefly.
- Please Turn Over.

Marginal Probability

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

- Suppose, the synthetic contingency table above represents the survey responses of BRAC University students regarding “Whether CGPA 3.5 should be added in the grading system or not?”

BRAC UNIVERSITY
Inspiring Excellence

Marginal Probability

- **Marginal Probability**

If the probability of **one** certain characteristic is computed, then it is known as the marginal probability. For example,

- **Example 01:** Calculate the probability of a randomly selected student is Freshmen.

Solution: P(Freshman) = 291/1218 = 0.23

We calculated it using the **classical approach.** There are 1218 students in total, and among them there are 291 freshmen. So, we just calculated the percentage of freshmen in the sample.

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Marginal Probability

- **Marginal Probability**

If the probability of only **one** certain characteristic is computed, then it is known as the marginal probability. For example,

- **Example 02**: Calculate the probability of a randomly selected student said Yes.

Solution: P(Yes) = 910/1218 = 0.74 (Ans)

- **Example 03:** Find P(N0).

Solution: P(No) = 308/1218 = 0.25 (Ans)

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Marginal Probability

- **Marginal Probability**

All marginal probabilities that can be computed from the contingency table 2.1 are listed below:

Find the probability that a randomly selected student :

1. Said Yes
2. Said No
3. Is Freshman
4. Is Sophomore
5. Is Junior
6. Is Senior
7. Is Advanced Senior

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Conditional Probability

- **Conditional Probability**

It is the probability that an event will occur given that another event has already occurred. If A and B are two events, then the conditional probability of A given B is written as P(A| B).

Reads as, Probability of A, given that B has already occurred. (Given that event B is already occurred, what is the probability of A?)

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

BRAC UNIVERSITY
Inspiring Excellence

Conditional Probability

- **Conditional Probability**
- **Example 01:** What is the probability of a randomly selected student is Advanced Senior, given that she said Yes?

Solution: Total number of students who said Yes = 910.

Advanced seniors who said Yes = 30.

So, P(Advanced Senior| Yes) = 30/910 = 0.03 (Ans)

- **Example 02:** Compute the probability of a student said No, given that he is a Sophomore.

Solution: Total number of sophomores = 246

Sophomores who said No = 36.

P(No| Sophomore) = 36/246 = 0.146 (Ans)

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Conditional Probability

- **Conditional Probability**
- **Example 03:** Calculate the probability that a randomly selected student is from the combined group of Seniors and Advanced Seniors, given that she said No.

Solution: Total number of students who said No = 308

Total number of students in the combined group of Seniors and Advanced Seniors who said No = (85+80) = 165.

So, P(Senior and Advanced Senior| No) = 165/308 = 0.53 (Ans)

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Addition Rule

- **Addition Rule**
- **Example 01:** Calculate the probability that a randomly selected student is from Freshmen or from Seniors.

Solution: P(freshman) = 291/1218 = 0.238

P(Senior) = 285/1218 = 0.234

So, P(Freshman or Senior) = 0.238+0.234 = 0.472 (Ans)

- **Example 02:** Calculate the probability that the student is Sophomore or Advanced Senior.

Solution: P(Sophomore) = 210/1218 = 0.172

P(Advanced Senior) = 30/1218 = 0.024

So, P(Sophomore or Advanced Senior) = 0.172+0.024 = 0.196(Ans)

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Addition Rule

- **Addition Rule**
- **Example 03:** Calculate the probability that a randomly selected student is from Freshmen or from the group said No.

Solution: P(Freshman) = 250/1218 = 0.238

P(said No but not a freshman) = (308-41)/1218 = 0.219

So, P(Freshman or No) = 0.238+0.219 = 0.457 (Ans)

Note: We did not use 308 in the nominator, rather we used 308-41. Since we already calculated the probability of being a freshman, so if we had to subtract that number for not causing overcalculation.

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Addition Rule

- **Addition Rule**
- **Example 04:** Calculate the probability that a randomly selected student is from Juniors or from the group said Yes.

Solution: P(Junior) = 220/1218 = 0.180

P(said Yes but not a Junior) = (910-220)/1218 = 0.566

So, P(Junior or Yes) = 0.180+0.566 = 0.746 (Ans)

Level	Yes	No	Total
Freshmen	250	41	291
Sophomores	210	36	246
Juniors	220	66	286
Seniors	200	85	285
Advanced Seniors	30	80	110
Total	910	308	1218

Table 2.1: Contingency Table

Finite Sample Transformation (FST)

Actual meaning of probability in FST

- It denotes the relative frequency of an outcome in a long term experiment.
- For example, if you want to calculate the probability of getting heads in a fair coin, you should coin the toss for a long time and record the outcome.
- If you toss it 500 times, you will see that you got head approximately 250 times.
- 250/500 = 0.5. This is the relative frequency of getting heads in a fair coin toss. This is also known as the probability of getting heads in a fair coin toss.

Finite Sample Transformation (FST)

Actual Meaning of Probability in FST

- Probability can be also explained using percentage. For example, if there are 3 red balls and 7 black balls in a basket (total 10 balls) then the probability of getting red balls is 0.3.
- Because, in the basket, there are 30% red balls. 30/100 = 0.3
- It means, if you keep picking balls with replacement for a long time (let, n= 500 again), you will see approximately 150 balls that you picked is red.

Finite Sample Transformation (FST)

Actual Meaning of Probability in FST

- So, when they say, "the probability of a certain Bangladeshi person likes Biryani is 0.98" what does that mean?
- It means that, if there are 100 people living in Bangladesh, then 98 of them likes Biryani.

Finite Sample Transformation (FST)

- **When to use Finite Sample Transformation?**

- When unable to use every method that is discussed before.
- When the sample size is not finite or not declared.
- It uses the formula pattern of computing marginal and conditional probabilities in contingency table to compute similar probabilities from infinite or not declared samples.

Finite Sample Transformation (FST)

- **Example 01:** The probability that a randomly selected student from a college is a senior is .20, and the joint probability that the student is a computer science major and a senior is .03. Find the conditional probability that a student selected at random is a computer science major given that the student is a senior.

Discussion

As you can see, you cannot apply the previous methods of calculating probability in this problem. Because it cannot be solved with discrete probability distributions, and cannot be solved by classical approach too. Because the sample size is not given, rather the probability is given. However, you can still solve such problems using Bayes' Theorem. But memorizing too much formulas will go heavy on you.

Finite Sample Transformation (FST)

- **Example 01:** The probability that a randomly selected student from a college is a senior is .20, and the joint probability that the student is a computer science major and a senior is .03. Find the conditional probability that a student selected at random is a computer science major given that the student is a senior.

Discussion

So, we will harness the power of relative frequency approach here. As mentioned before, probability is nothing but the relative frequency, and relative frequency is just percentages. Here's a reminder. We know that the probability of getting a head in a fair coin toss is 0.5. Now, if you toss a coin for a long time (like 1000 times), you'll see that you got approximately 500 heads.

Finite Sample Transformation (FST)

- **Example 01:** The probability that a randomly selected student from a college is a senior is .20, and the joint probability that the student is a computer science major and a senior is .03. Find the conditional probability that a student selected at random is a computer science major given that the student is a senior.

Discussion

Here, percentage of heads = (500/1000 * 100)% = 50%

Relative frequency of heads = 500/1000 = 0.5

And we know, theoretical probability of getting heads in a fair coin toss is 0.5

So, probability is nothing but the relative frequency of a repetitive event, and the percentage. Hence, we will try to transform the entire questions into percentage statements by making the sample size finite using the given hypothetical relative frequency.

Finite Sample Transformation (FST)

- **Example 01:** The probability that a randomly selected student from a college is a senior is .20, and the joint probability that the student is a computer science major and a senior is .03. Find the conditional probability that a student selected at random is a computer science major given that the student is a senior.

Solution: Let, total students in that college is 1000 (finite transformation)

Given that, The probability that a randomly selected student from a college is a senior is .20.

So, number of senior students in that college (1000 * 0.20) = 200

Given that, the joint probability that the student is a computer science major and a senior is .03.

So, number of students who are seniors and CS major = (1000 * 0.03) = 30

Hence, P(CS and Senior| Senior) = 30/200 = 0.15 (Ans)

Finite Sample Transformation (FST)

- **Example 02**: The weather forecast predicts a 70% chance of rain tomorrow. Based on past data, when the forecast predicts rain, it actually rains 80% of the time, and when the forecast predicts no rain, it rains 10% of the time. Given that it rained, what is the probability that the forecast predicted rain?

Solution: Let, past record on 1000 days (finite sample transformation)

In past 1000 days where weather was like tomorrow's prediction (1000*0.7)=700 days were predicted to be rained.

Among these 700 days, (700*0.8) = 560 days were really rained.

In past 1000 days where weather was like tomorrow's prediction, (1000-700)= 300 days were predicted not to be rained.

Among these 300 days, (300*0.1) = 30 days were really rained.

So, P(Forecast predicted rain| It rained) = 560/(560+30) = 0.949 (Ans)

BRAC UNIVERSITY
Inspiring Excellence

Finite Sample Transformation (FST)

- **Example 03:** Machine A,B, and C respectively yields 50%, 30%, and 20% of the output. The defect rates for these machines are respectively 2%, 3% and 5%. If a randomly selected product is defective, compute the probability that it is from machine B or C.

Solution: Let, total products = 100 (finite sample transformation)

Machine A,B, and C produces respectively 50, 30, and 20 products.

Machine A,B, and C produces respectively 1, 0.9, and 1 defective products.

Total defective products = 1+0.9+1 = 2.9

So, P(B or C| defective) = 0.9+1/2.9 = 0.655 (Ans)

Mutually Exclusive Events

- Events that cannot occur together are known as mutually exclusive events. For example, we cannot get a head and a tail at the same time from a toss of a fair coin. Getting head and getting tail are mutually exclusive events in this context.
- Event A and Event B is mutually exclusive if, P $(A \cap B)$= P(A and B) = 0

Mutually Exclusive Events

Example 1: A box contains a total of 100 CDs that were manufactured on two machines. Of them, 60 were manufactured on Machine I. Of the total CDs, 15 are defective. Of the 60 CDs that were manufactured on Machine I, 9 are defective. Let D be the event that a randomly selected CD is defective, and let A be the event that a randomly selected CD was manufactured on Machine I. Are events D and A mutually exclusive?

Solution: Given, Event D = A randomly CD is defective.

Event A = A randomly selected CD was manufactured on Machine I

So, If P(Randomly selected CD is defective **and** came from Machine I) = P(D and A) = 0, then the events are mutually exclusive events. Otherwise, not.

P(D and A) = 9/100 = 0.09

Hence, the events are not mutually exclusive.

Independent and Dependent Events

- If the occurrence of one event does not affect the occurrence of another event, then the two events are **independent events.** Two events A and B are independent if, P(A| B) = P(A), or, P(B| A) = P(B)

Example 1: A box contains a total of 100 CDs that were manufactured on two machines. Of them, 60 were manufactured on Machine I. Of the total CDs, 15 are defective. Of the 60 CDs that were manufactured on Machine I, 9 are defective. Let D be the event that a randomly selected CD is defective, and let A be the event that a randomly selected CD was manufactured on Machine I. Are events D and A independent?

Solution: P(D) = 15/150 = 0.15

P(D| A) = 9/60 = 0.15

Since P(D) = P(D| A). The events are independent. (Ans)

Counting Rule

- We calculate the total number of outcomes in an experiment using counting rule.
- If an experiment is of four steps, and each steps have respectively w, x, y, z outcomes, then total outcomes for the experiment = w * x * y * z.

Example 01: Compute the total number of outcomes if a coin is tossed 5 times.

Solution: Each toss have 2 outcomes.

So, total outcome = $2^5 = 32\ (Ans)$

Counting Rule

- We calculate the total number of outcomes in an experiment using counting rule.
- If an experiment is of four steps, and each steps have respectively w, x, y, z outcomes, then total outcomes for the experiment = w * x * y * z.

Example 02: A random number generator program runs in a "for loop" 20 times and prints a number from the list [2,5,7]. What is the total outcome for the experiment?

Solution: Each loop can print any of the 3 numbers in the list. So, there are 3 outcomes.

The loop runs 20 times.

So, total outcome = 3^{20} (Ans)

Chapter 03: Linear Relationship

KAZI SAKIB HASAN (24341237)

COMPUTER SCIENCE PROGRAM, BRAC UNIVERSITY

kazi.sakib.hasan@g.bracu.ac.bd

Brief Contents

CORRELATION

- PEARSON'S CORRELATION COEFFICIENT
- COEFFICIENT OF DETERMINATION
- SCATTER PLOT

SIMPLE LINEAR REGRESSION

- LEAST SQUARE REGRESSION

Correlation

- Sometimes we need to find out whether two variables have any linear relationship (increasing-decreasing) with each other or not. This relationship between two variables is known as **correlation.** Among them there is one **dependent variable**, and the other one is **independent variable**.
- Some examples on correlation are given below.

- Education level and income: **Income amount** usually increases alongside the **education level** of a person.
- Code complexity and bug count: **The more complex a program** is, the more **buggy** it can be.
- Server load and response time: When the **load on server** is too much, its **response time** decreases.
- Sunlight exposure and plant growth: Increasing of **sunlight exposure** can increase **plant growth**.
- Ohm's law: As the **voltage** across the resistor increases, the **current** through the resistor also increases.

BRAC UNIVERSITY
Inspiring Excellence

Correlation Coefficient

- Correlation coefficient describes the linear relationship between two variables. Its value ranges from -1 to 1. Correlation coefficient is denoted as $\boldsymbol{\rho}$ in population data, and in sample data it is denoted as $\boldsymbol{r}$. Our discussion is confined within the sample data only.
- We take decisions on linear relationship based on the value of r and coefficient interval.

Coefficient Interval	Correlation	Coefficient Interval
0.00 – 0.199	Very weak	0.00 – (-0.199)
0.20 – 0.399	Weak	-0.2 – (-0.399)
0.40 – 0.599	Moderate	-0.40 – (-0.599)
0.60 – 0.799	Strong	- 0.60 – (-0.799)
0.80 – 1.00	Very Strong	-0.80 – (-1.00)

Table 3.1: Coefficient intervals and their corresponding strength of correlation.

Correlation Coefficient

- The formula of calculating correlation coefficient r is the same for population data and sample data. Note that, the correlation coefficient calculation formula that is given below is known as Pearson's correlation coefficient.
- Formula: $r = \frac{\sum x_i y_i - n.\bar{x}.\bar{y}}{\sqrt{(\sum x_i^2 - n.\bar{x}^2)\,(\sum y_i^2 - n.\bar{y}^2)}}$

Here, x_i = Every data point in x variable (independent variable)

y_i = Every data point in y variable (dependent variable)

$\bar{x}$ = Arithmetic mean, or average of x variable

$\bar{y}$ = Arithmetic mean, or average of y variable

$\sum x_i^2$ = Summation of every squared data point in x variable

$\sum y_i^2$ = Summation of every squared data point in y variable

Correlation Coefficient

- **Correlation does not always indicate causation**

- A strong linear relationship between variables does not indicate that the independent variable is causing the dependent variable. To exemplify, assume that the correlation coefficient between numbers of ice-cream sold and numbers of people drowned in the swimming pools is 0.75, which denotes a strong relationship.
- It does not indicate that people who were consuming ice-cream ended up drowning in the pools. In most cases, there is at least one **lurking variable** that manipulates the dependent variable to act in such way.
- In this example, the **lurking variable** can be a natural season. People relatively consume ice-cream in the summer more than the other seasons. At the same time, people tend to swim more in the summer. Hence, there will be more drowning as well.
- So, the correlation coefficient interpretation should not be like "Consuming ice-cream makes people drown", rather the coefficient just denotes that both variables are increasing with each other. Now, if that increasing is concerning, then researchers find out the lurking variables behind all of these to exterminate the danger.

Correlation Coefficient

- **Correlation rarely indicates causation**

- There are some variables that will always perfectly correlate with each other. These variables are usually from mathematical expressions.
- For example, the relationship between voltage and electricity in Ohm's law, the relationship between mass and force when the acceleration is constant in Newton's law etc. In such cases of mathematical expressions, indeed the independent variables affect the dependent variable so perfectly that the value of r becomes 1.00. In these cases, correlation denotes causation.
- But those relationships that are not pure mathematical (as mentioned in page 3), the value of r will be rarely1.00 in that cases. Even if it becomes 1.00, still that does not indicate causation. In these cases, correlation does not denote causation.

BRAC UNIVERSITY
Inspiring Excellence

Correlation Coefficient

- **What causes what?**

- Sometimes we cannot understand what is the independent variable in an experiment. For example, assume several people are selected from a city and it is found that their depression level and addiction to drugs strongly correlates with each other.
- In this case, we cannot tell which is the independent variable here. Because maybe they are depressed, and that's why they take drug pretty much. Or maybe they do drug too much, and hence they are depressed. This is also a reason, why we say that correlate does not indicate causation.

Correlation Coefficient

- **Example 01:** 12 software engineers from a certain company is randomly selected and tasked to write the backend codes for a complex website. After the task is finished, the bug counts and Lines of Code (LoC) are reported. Calculate the correlation coefficient and comment on this.

  LoC = [201, 188, 199, 205, 243, 155, 238, 173, 160, 221, 250, 190]

  Bugs = [5, 3, 4, 4, 7, 3, 6, 6, 5, 2, 7, 3]

Solution: Number of bugs is generally dependent on lines of code. Complex and lengthy codes produces more bugs in the code. Therefore, in this context, LoC is the independent variable (x) and Bugs is the dependent variable (y).

Correlation Coefficient

Solution: Given,

LoC = [201, 188, 199, 205, 243, 155, 238, 173, 160, 221, 250, 190]

Bugs = [5, 3, 4, 4, 7, 3, 6, 6, 5, 2, 7, 3]

Here, $\bar{x} = \frac{201+188+199+205+243+155+238+173+160+221+2}{12} = 201.91$

$\bar{y} = \frac{5+3+4+4+7+3+6+6+5+2+7+3}{12} = 4.58$

$\sum x_i y_i = (201 * 5) + (188 * 3) + (199 * 4) + (205 * 4) + (243 * 7) + \cdots + (190 * 3) = 11379$

$\sum x_i^2 \ = (201)^2 + (188)^2 + (199)^2 + (205)^2 + \cdots + (190)^2 = 500059$

$\sum y_i^2 = 5^2 + 3^2 + 4^2 + 4^2 + \cdots + 3^2 = 283$

Now, $r = \frac{11379 - 12 * 201.91 * 4.58}{\sqrt{(500059 - 12 * (201.91)^2)\ (283 - 12 * (4.58)^2)}} = 0.484$ (Ans)

The correlation coefficient r is 0.484. Therefore, a **moderate positive correlation** exists between the variables “number of bugs” and “lines of codes”. (Ans)

Coefficient of Determination

- Coefficient of determination tells us the percentage of variation in dependent variable due to the variation of independent variable. You can replace the word "variation" with "changes".
- Coefficient of determination is calculate by squaring the value of correlation coefficient, and so it is denoted as r^2. It is also known as **goodness of fit.**
- If the coefficient of determination for two variables is 0.7225, it means that (0.7225)*100% = 72.25% data point in the dependent variable (y) can be explained by the variation of data points in the independent variable (x). Again, do not forget that this variation does not denote to causation every time.

Coefficient of Determination

- **Example 02:** The correlation coefficient between lines of codes (LoC) and number of bugs is 0.484. Compute the coefficient of determination and interpret the result.

Solution: Given, $r = 0.484$. So, coefficient of determination $r^2 = (0.484)^2 = 0.234$ (Ans)

$0.234 * 100\% = 23.4\%$

Therefore, 23.4% variations in the number of bugs (dependent variable) can be explained by the variations in the lines of codes (independent variable). The rest are might be significantly affected by lurking variables like coding experience, processing problems, lack of planned designing etc. (Ans)

Coefficient of Determination

- **Example 03:** A sample of 1020 students from a high school is selected and it is found that the correlation coefficient between total hours of playing games in a day and the stress level calculated with PSS-10 scale is 0.83. What is your thoughts regarding this?

Solution: The value of r is 0.83 which indicates a strong positive correlation between amount of playing games and stress level, which is quite concerning. There can be three cases.

1. Video games are making the teens stressed.
2. Stress is making the teens to play video games to obtain relief from it.
3. There are lurking variables that are controlling the teens to play video games, and be stressed at the same time.

Scatter Plot

- Scatter plot/ scatter diagram helps us to visualize the linear relationship between two variables without calculating the correlation coefficient. It is a fast inferential approach to get a visual on the correlation between variables.

- **How to draw a scatter plot?**

- Its simple. First of all, place every values (data points) of independent variable on the x-axis. Then, place every values (data points) of dependent variable on the y-axis.
- In short, plot x dataset values in x-axis and y dataset values in y-axis.

Scatter Plot

- **Example 01:** The datasets on sold ice-creams and number of people went for swimming in each day is given. Draw a scatter diagram.

  Sold ice-creams = [45, 33, 51, 60, 76, 29, 65, 57, 59, 64]

  People swam = [50, 37, 56, 63, 77, 26, 66, 62, 64, 59]

Solution: Here, let, x = Sold ice-creams and y = People swam

Now, we get the scatter diagram by plotting the values of x in x-axis, and values of y in y-axis.

(Please Turn Over)

Scatter Plot

- **Example 01:** The datasets on sold ice-creams and number of people went for swimming in each day is given. Draw a scatter diagram.

  Sold ice-creams = [45, 33, 51, 60, 76, 29, 65, 57, 59, 64]

  People swam = [50, 37, 56, 63, 77, 26, 66, 62, 64, 59]

Solution:

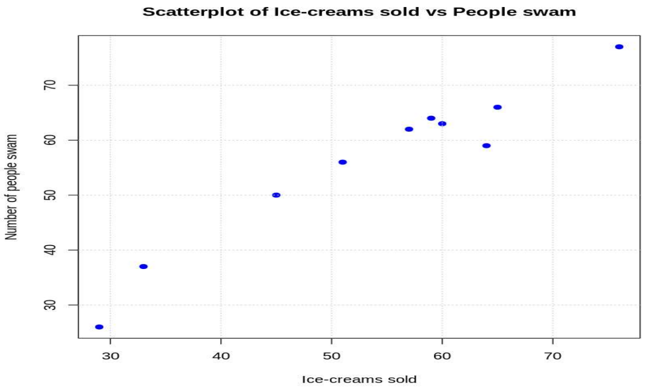


Fig 3.1: Scatterplot

The scatterplot in fig 3.1 is drawn using R. As you can see, the dots are increasing. It represents that the variables are positively correlated with each other.

Scatter Plot

- **Example 01:** The datasets on sold ice-creams and number of people went for swimming in each day is given. Draw a scatter diagram.

  Sold ice-creams = [45, 33, 51, 60, 76, 29, 65, 57, 59, 64]

  People swam = [50, 37, 56, 63, 77, 26, 66, 62, 64, 59]

Solution:

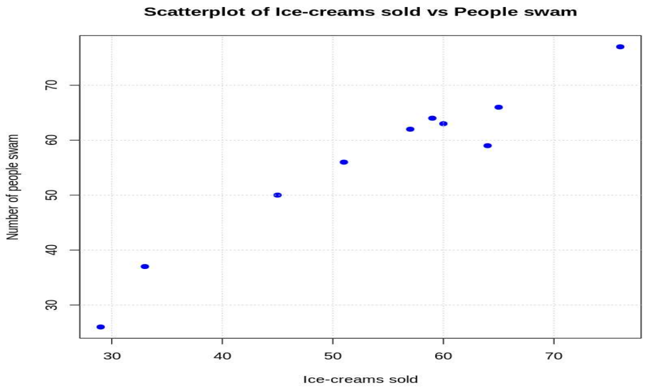


Fig 3.1: Scatterplot

Note that, programming languages like R and Python, simply plots y-values on the y-axis ignoring the x-values. Because the x-values falls on the horizontal line of x-axis that does contribute to the visualization of linear relationships.

Simple Linear Regression

- When two variables have a good strength in correlation, we can build a predictive model to predict a certain value for dependent variable, based on the corresponding value of independent variable.
- This type of task is done by using **Linear Regression** models.
- When a linear regression model is built using two variables only, we call it **Simple Linear Regression (SLR)** model. Conversely, linear regression model with multiple dependent variables and one dependent variables are known as **Multiple Linear Regression** model.
- The equation of simple linear regression model came from the straight line equation $y = mx + c$.
- Our discussion will be confined within **Simple Linear Regression** only.

Simple Linear Regression

- **What we do in Linear Regression?**

- We make a scatterplot of the dependent and independent variable, and then draw straight line through the points of scatterplot in such way that every point in that plot have the minimum distance from the line.
- The line is called as **best fitting line** **or** **the trend line.**
- The distance from the line to the points are called as **residuals.**
- As the best fitting line is a straight line equation that can be represented as $y = mx + c$, so, by inputting the value for x, we can estimate the value of y. This estimate might not be fully accurate due to the residual, but the estimation still can give a major insight.

Simple Linear Regression

- Recognize the function concept : y = f(x). Here, the values of y is dependent on the values of x.
- Assume, f(x) = 3x + 5
- So, for each values of x, y will increment.
- Therefore, we can say that there is a **relationship** between x and y variables, as their values are being changed with each other.
- For the previous function, y = 3x+5. By plugging in random x values, we get the values for y.
- x = [1, 2, 3, 4, 5] (independent variable)
- y = [8, 11, 14, 17, 20] (dependent variable)
- In our real life, we have access to the datasets of independent and dependent variables only, we do not get the functions directly. We need to derive the function for predictive modelling from the independent and dependent variables. We derive the function using **Simple Linear Regression (SLR).**

Simple Linear Regression

- Topics to be discussed:
- Least Squares Regression
- Slope
- x-intercept
- y-intercept
- Residuals
- Extrapolation

Every topic will be explained using a mathematical example for ease.

BRAC UNIVERSITY

Inspiring Excellence

Least Squares Regression

- Least Squares Regression is a method that produces the equation for the best fitting line. The equation is: $\hat{y} = \beta_1 x + \beta_0$
- This is the equivalent equation to $y = mx + c$
- In the equation,

  $\hat{y}$ = y = predicted value for dependent variable

  β_1 = m = slope.

  x = Independent variable

  β_0 = c = y-intercept.
- If we can calculate the values for β_1 and β_0, then we can use the linear regression model for prediction.

Least Squares Regression

- $\beta_1 = \frac{S_y}{S_x} * R$
- $\beta_0 = \bar{y} - \beta_1 \bar{x}$
- In the equations stated above,

  S_y = Standard deviation of y dataset (dependent variable)

  S_x = Standard deviation of x dataset (independent variable)

  R = Pearson correlation coefficient

  $\bar{y}$ = Mean of y dataset (dependent variable)

  $\bar{x}$ = Mean of x dataset (independent variable)
- Calculating the value for x-intercept is also often required.

  x-intercept = $-\frac{\beta_0}{\beta_1}$

Least Squares Regression

- Some synonyms are important to learn in least squares regression.

Slope	y-intercept	x-intercept	Independent variable	Dependent variable
Regression coefficient	Intercept	Zero of the function	Explanatory variable	Response variable
Coefficient of x	y value when x = 0	x-value when y = 0	Predictor variable	Outcome variable
Rate of change	Constant term	Root value of the equation	Input variable	Target variable
Gradient			Regressor	Criterion

Table 3.2: Synonyms for terminologies used in Least Squares Regression

Least Squares Regression

- **Understanding independent and dependent variables**

- Understanding which variable is dependent and which one is independent is often the first step before performing linear regression. We need to infer this using our general knowledge.
- Variables that can cause the other variable to act in its way, is the independent variable. Conversely, variable that can be affected by the other remaining variable is the dependent variable. Even though correlation does not denote causation every time, but sometimes it does, and hence we assume in the beginning that independent variable manipulates the dependent variable.
- For example, assume that girls of different ages are selected and asked about the estimation of how many times they are touched without their consent in the public transports. What can be the independent and dependent variables here? We can infer from our common sense that kids are not sexually harassed in public transports, but the teenage and youth girls are harassed often. So, age **might be** a variable to manipulate harassment rates. Therefore, in this context, age is the independent variable and harassment rates are the dependent variable.

BRAC UNIVERSITY
Inspiring Excellence

Least Squares Regression

- **Example 01:** The natural logarithmic transformation of array sizes and program runtimes in seconds are given. Each program runs in linear time complexity O(n).

  array_sizes = [1.60, 4.60, 5.70, 6.55, 7.31, 7.69]

  program_runtimes = [0.49, 3.56, 4.01, 6.56, 8.58, 9.25]

A. Fit an equation using least squares regression methods.

B. Interpret regression coefficient, constant term, and x-intercept.

C. Create the scatterplot and fit the straight line.

D. Predict estimated runtime for an array with natural log transformed size 6.68.

E. Calculate residual for natural log transformed array size 5.70.

F. Find R and R^2 and interpret both.

G. If the natural log transformation of an array size is 8.29, should you predict its program runtime from the best fit equation?

Least Squares Regression

- **Solution:** First of all, we need to find out what the independent and dependent variable is by ourselves, since its not mentioned in the question. In computer programs, the program runtimes are dependent on array sizes. Its common sense that the larger the array size is, the slower the processing time is. So, program runtimes are supposed to be increased with the array size. So, in this context,

Array size is the explanatory variable x (independent variable).

Program runtime is the response variable y (dependent variable).

Least Squares Regression

- **Solution A:** The equation for the best fitting line according to least squares regression, $\hat{y} = \beta_1 x + \beta_0$

Here, $\beta_1 = \frac{S_y}{S_x} * R$ and $\beta_0 = \bar{y} - \beta_1 \bar{x}$

Calculating the values for S_y, S_x, and R, we get,

$\beta_1 = \frac{3.337}{2.245} * 0.95 = 1.412$

Calculating the values for $\bar{y}$ and $\bar{x}$, we get,

$\beta_0 = \bar{y} - \beta_1 \bar{x} = 5.408 - 1.412 * 5.574 = -2.462$

So, the equation for the least squares regression line is: $\hat{y} = 1.412x - 2.462$ (Ans)

Least Squares Regression

- **Solution B:** From Solution A, regression coefficient, $\beta_1 = 1.412$

Constant term, $\beta_0 = -2.242$

So, x-intercept = $-\frac{-2.242}{1.412} = 1.96$

Regression coefficient $\boldsymbol{\beta_1}$ interpretation: For each unit increases or decreases in $\boldsymbol{x}$, $\hat{\boldsymbol{y}}$ correspondly increases or decreases by β_1 unit. In the context of array size and runtimes problems, for every 1 unit increase in array size, the runtimes increase by 1.412 seconds.

Constant term $\boldsymbol{\beta_0}$ interpretation: The β_0 is the value of $\hat{y}$ when $\boldsymbol{x = 0}$. The interpretation for constant term does not always have practicality. For example, in array size and runtime context, when the array size is 0, the program runtime is

-2.242s, which is not possible in real life.

X-intercept interpretation: It is the value of x, when y = 0. It also does not have practicality always. For example, in our context, the interpretation is that it takes 0 seconds to process an array with size 1.96. In reality, nothing can be processed within 0 seconds.

Least Squares Regression

- **Solution C:** The scatterplot with the least squares regression line is shown below:

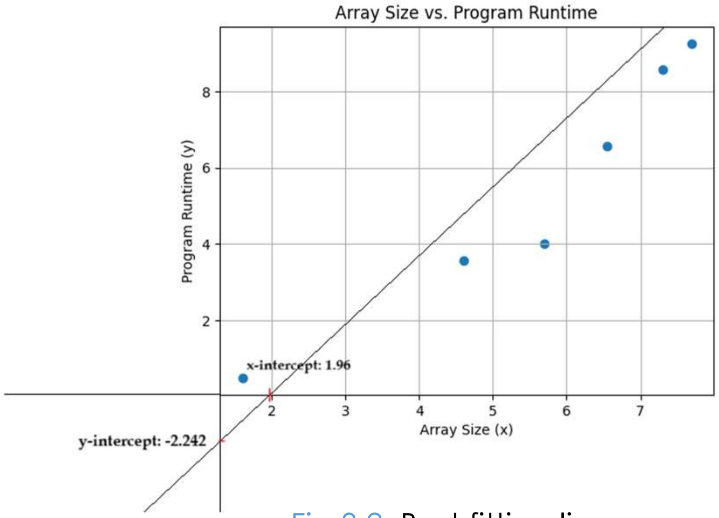


Fig 3.2: Best fitting line

Least Squares Regression

- **Solution D:** Given, x = 6.68

Best fitting line equation: $\hat{y} = 1.412x - 2.462$

So, $\hat{y}$ = $1.412 * 6.68 - 2.462 = 6.97$s

Therefore, estimated runtime for an array with natural log transformed size 6.68 is 6.97s (Ans)

- **Solution E:** Given, the corresponding runtime for array size 5.70 is 4.01s. So, observed value, $y_i = 4.01$s

Predicted value $\hat{y}$ = $1.412 * 5.70 - 2.462 = 5.58$s [Using best fitting line equation]

Now, residual = y_i - $\hat{y}$ = 4.01 - 5.58 = -1.57s (Ans)

Least Squares Regression

- **Solution E:** We already calculated the value for Pearson correlation coefficient R in Solution A. R = 0.95.

So, coefficient of determination $R^2 = (0.95)^2 = 0.9025$

$0.9025 * 100\% = 90.25\%$

Interpretation: A very strong positive correlation with a coefficient of 0.95 exists between the variables array size and program runtimes.

90.25% variations in the program runtimes can be explained by the variations in the array sizes.

Least Squares Regression

- **Solution F:** The data point 8.29 is greater than the largest value in the dataset of independent variable "array size". So, predicting the runtime for this array size will be an **extrapolation.**

The linear relationship is observed within a certain range. Its uncertain that whether the relationship is still the same outside the range or not, so if the natural log transformation of an array size is 8.29 ($x = 8.29$), predicting its program runtime ($\hat{y}$) from the best fit equation is not a good choice. Sometimes the extrapolation might provide a good estimation, but we will not know whether the estimation was quite accurate or not.

Note: Even if the given value for $\boldsymbol{x}$ was smaller than the smallest value in the dataset, then predicting $\hat{\boldsymbol{y}}$ could still result in extrapolation.

In summary, do not predict $\hat{\boldsymbol{y}}$ if its corresponding $\boldsymbol{x}$ value is smaller than the smallest value in the independent dataset, or greater than the greatest value in the independent dataset.

However, if you know that the entire data beyond your sample follows the linear relationship, you can do extrapolation to predict $\hat{\boldsymbol{y}}$ values beyond the limit of independent dataset.

Chapter 04: Random Variables

KAZI SAKIB HASAN (24341237)

COMPUTER SCIENCE PROGRAM, BRAC UNIVERSITY

kazi.sakib.hasan@g.bracu.ac.bd

Brief Contents

DISCRETE RANDOM VARIABLES (DRV)

- PROBABILITY DISTRIBUTION OF DRV
- MEAN OF PROBABILITY DISTRIBUTION OF DRV
- SD OF PROBABILITY DISTRIBUTION OF DRV
- FAIR GAME
- DISCRETE PROBABILITY DISTRIBUTIONS

CONTINUOUS RANDOM VARIABLES (CRV)

- PDF AND CDF
- CONTINUOUS PROBABILITY DISTRIBUTION: EXPONENTIAL DISTRIBUTION

Random Variables

- A random variable (usually denoted as X) represents the outcome of an experiment/ random phenomenon/ random process. For example,
- If a six-sided die is rolled, there can be six outcomes. S = {1,2,3,4,5,6}. If the outcome is 3, then random variable x = 3. Similarly, for each experiment, x takes on the value of the outcome.
- Two types of Random Variables :

1. Discrete Random Variables
2. Continuous Random Variables

We will discuss both in this chapter.

Discrete Random Variables (DRV)

- Discrete Random Variables are those variables that carry discrete data points. For example, let X be the discrete random variable of :

- Number of failures in a circuit board test.
- Number of errors in lines of codes.
- Number of cells killed by an antibiotic.
- Moneys earned from a lotto game.

All of the data points in these variables are discrete. For example, if X is the discrete random variable containing number of failures in a circuit board test, then x might be [1,2,3,4,….] in a random experiment, where $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ and $X =$ "Number of failures" dataset, or variable that can take on any non-negative x_i values.

Probability Distribution of DRV

- Assume, the developer of a certain website collected the data for various users regarding their number of login attempts with wrong credentials before the first success. The probability distribution is shown in table 4.1. Probability distribution was also discussed in pre-requisites.

# of attempts (X)	Frequency (f)	Relative Frequency (f_i)	Probability P(x)
$x_1 = 0$	155	155/500 = 0.31	0.31
$x_2 = 1$	100	100/500 = 0.2	0.2
$x_3 = 2$	102	102/500 = 0.204	0.204
$x_4 = 3$	89	89/500 = 0.178	0.178
$x_5 = 4$	54	54/500 = 0.108	0.108
	$\sum f = 500$	$\sum f_i = 1.00$	$\sum P(x) = 1.00$

Table 4.1: Frequency distribution, relative frequency distribution, and probability distribution for number of wrong login attempts by users.

Probability Distribution of DRV

- Two characteristics of probability distribution:

1. For each value of x, $0 \leq \mathrm{P(x)} \leq 1$
2. $\sum P(x) = 1$

Example 01: What is the probability that a randomly selected user from the sample attempted more than 2 times? (From table 4.1)

Solution: $P(x > 2) \ = \ P(x = 3) \ + \ P(x = 4) \ = \ 0.178 + 0.108 \ = \ 0.286$ (Ans)

Mean of Probability Distribution of DRV

- The formula to calculate the mean of probability distribution of discrete random variable is given below:

$$Expected\ value, u = E(x) = \sum xP(x)$$

There is another formula to calculate the mean. We can use it if the frequency distribution is available to us like table 4.1.

$$Mean = \frac{\sum xf}{\sum f}$$

The expected value (mean) for the probability distribution in table 4.1 is calculated in the next page.

BRAC UNIVERSITY
Inspiring Excellence

Mean of Probability Distribution of DRV

- Table 4.2 only illustrates the probability distribution for login attempts of the users showed in table 4.1

# of attempts (X)	Probability P(x)
$x_1 = 0$	0.31
$x_2 = 1$	0.2
$x_3 = 2$	0.204
$x_4 = 3$	0.178
$x_5 = 4$	0.108
	Total = 1.00

Table 4.2: Probability distribution for number of login attempts by users

From table 4.2,

$$\begin{aligned} \mathrm{E(x)} &= (0 * 3.1) + (1 * 0.2) + (2 * 0.204) + (3 * 0.178) + (4 * 0.108) \\ &= 0 + 0.2 + 0.408 + 0.534 + 0.432 \\ &= 1.574 \text{ (Ans)} \end{aligned}$$

Practical Interpretation: On average, if observed for long periods, it can be noticed that the users attempt 1.574 times to log into the website with wrong credentials.

Mean of Probability Distribution of DRV

- Table 4.3 only illustrates the frequency distribution for login attempts of the users showed in table 4.1

# of attempts (X)	Frequency (f)
$x_1 = 0$	155
$x_2 = 1$	100
$x_3 = 2$	102
$x_4 = 3$	89
$x_5 = 4$	54
	$\sum f$ = 500

Table 4.3: Frequency distribution for number of login attempts by users

From table 4.3,

$\sum xf$ = (0 * 155) + (1 * 100) + (2 * 102) + (3 * 89) + (4 * 54)
= 787
$\sum f$ = 500

So, E(x) = 787/500 = 1.574

Practical Interpretation: On average, if observed for long periods, it can be noticed that the users attempt 1.574 times to log into the website with wrong credentials.

SD of Probability Distribution of DRV

- Standard deviation (σ) of a discrete random variable measures the spread of its probability distribution.
- Higher σ denotes that x can take values over a large range about the mean.
- Lower σ denotes that x can take values over a small range about the mean.
- Standard deviation of probability distribution for discrete random variables can be calculated using two formulas. They are shown in the next page.

SD of Probability Distribution of DRV

- Basic formula: $\sigma = \sqrt{\sum[(x-\mu)^2 . P(x)]}$
- Shortcut formula: $\sigma = \sqrt{\sum x^2 P(x) - \mu^2}$

You can calculate the standard deviation using the frequency distribution too, but that would be too rigorous to compute using pen, papers, and usual calculators. Programming software would require to do so. Hence, its better to use either the basic formula or the shortcut formula.

In the next page, the standard deviation for the probability distribution shown in the table 4.2 is calculated using the both formulas.

SD of Probability Distribution of DRV

- Basic formula: $\sigma = \sqrt{\sum[(x-\mu)^2.P(x)]}$
- Shortcut formula: $\sigma = \sqrt{\sum x^2P(x) - \mu^2}$, $\mu = \mathrm{E}(\mathrm{x}) = 1.574$

# of attempts (X)	Probability P(x)	$(x-\mu)^2.P(x)$	$x^2P(x)$
$x_1 = 0$	0.31	$(0-1.574)^2.0.31$	0
$x_2 = 1$	0.2	$(1-1.574)^2.0.2$	$1^2.0.2$
$x_3 = 2$	0.204	$(2-1.574)^2.0.204$	$2^2.0.204$
$x_4 = 3$	0.178	$(3-1.574)^2.0.178$	$3^2.0.178$
$x_5 = 4$	0.108	$(4-1.574)^2.0.108$	$4^2.0.108$
	Total = 1.00	$\Sigma[(x-\mu)^2.P(x)] =$ 1.865	$\sum x^2P(x) = 4.346$

Table 4.4: Probability distribution and variances for number of login attempts by users

Using basic formula,
$\sigma = \sqrt{1.865} = 1.366\ (Ans)$

Using shortcut formula,
$\sigma = \sqrt{4.346 - (1.574)^2} = 1.366\ (Ans)$

Practical Interpretation: If the server is observed over long periods, the expected login attempt to enter the website with wrong credential is within the interval: (1.574 ± 1.366)

Fair Game

- **Example 01:** There are 38 slots in American roulette games: 18 red, 12 black, and 2 green. Gamblers (players) can place bets on red or black. If the ball stops on their color, the money of players is doubled. Otherwise, they lose their money. Suppose, Ash bet $2 on red.

A. Create a probability distribution model from the scenario.

B. Calculate expected value and standard deviation. Interpret the results.

C. Should Ash play the game regularly? Is it a fair game?

BRAC UNIVERSITY

Inspiring Excellence

Fair Game

- **Solution A:** The probability distribution for the scenario is given below:

Outcome	Profit (X)	Probability P(x)
Red	2	18/38 = 0.47
Black or Green	-2	20/38 = 0.52

Table 4.1.1: Probability distribution for the scenario

BRAC UNIVERSITY
Inspiring Excellence

Fair Game

- **Solution B:** The probability distribution for the scenario is given below:

Outcome	Profit (X)	Probability P(x)	xP(x)	$x^2P(x)$
Red	2	$18/38 = 0.47$	2*0.47	$2^2 * 0.47$
Black or Green	-2	$20/38 = 0.52$	-2*0.52	$(-2)^2 * 0.52$
			E(x)= -0.1	$\sum x^2P(x) = 3.96$

Table 4.1.2: Probability distribution, and variances for the scenario

E(x) = -0.1 (Ans)

SD = $\sqrt{3.96 - (-0.1)^2} = 1.98$ (Ans)

Interpretation for E(x): The players are expected to lose -0.1 dollars per match.

Interpretation for SD: The interval for the expected value is (-0.1 ± 1.98). If the game is played for a long days, we expect that the winnings from that game will be between -2.08$ to 2.08$.

Fair Game

- **Solution C:** The expected value is -0.1$. It means that the players are expected to lose -0.1 dollars per match. So, in the long run, the gamblers with keep loosing their money and the organizers will continue earning their money. This is how they make profit and run their business. So, Ash should not continue playing the game. It is not a fair game. A fair game is such games where the expected value for both the players and the organizers is 0$.

Discrete Probability Distributions

- **Four popular discrete probability distributions are :**

1. Binomial Distribution
2. Hypergeometric Distribution
3. Geometric Distribution
4. Poisson Distribution

We discussed about them in Chapter 02.

Continuous Random Variables (CRV)

- Continuous random variables carry continuous data points and these points are infinite within a certain range.
- Some continuous random variables:

1. Time taken to transfer a file over the network.
2. The percentage of CPU utilization at any given moment.
3. The frequency of any signal.
4. The output voltage of any electric circuit.
5. The size of human cells in any measurement unit.

Continuous Random Variables (CRV)

- The times taken (minutes) to share 5 GB sized files via a network are given in the table below.

Times taken (x)	f	$\boldsymbol{f_i}$
62 – 63	460	0.114
63-64	750	0.186
64 – 65	970	0.241
65 – 66	760	0.189
66 – 67	640	0.159
67 - 68	440	0.109
	$\sum f = 4020$	$\Sigma f_i = 1.00$

Table 4.2.1: Frequency and relative frequency distributions of times taken to transfer 5 GB files.

BRAC UNIVERSITY
Inspiring Excellence

Continuous Random Variables (CRV)

- The times taken (minutes) to share 5 GB sized files via a network are given in the table below.

Times taken (x)	f	f_i
62 – 63	460	0.114
63-64	750	0.186
64 – 65	970	0.241
65 – 66	760	0.189
66 – 67	640	0.159
67 - 68	440	0.109
	$\sum f = 4020$	$\Sigma f_i = 1.00$

Table 4.2.1: Frequency and relative frequency distributions of times taken to transfer 5 GB files.

A Fact: In CRV, you can calculate the probability of a randomly selected value falls under a certain interval, but cannot calculate the probability of getting that single value. It happens because the data is in the interval (bins) form, not in the single classed form. (Unlike DRV)

So, $\boldsymbol{P(a < x < b)}$ **is a value between 0 to 1.**
But, $\mathrm{P(a) = P(b) = 0}$.
Also, $\boldsymbol{P(a < x < b) = P(a \le x \le b) = P(a < x \le b) = P(a \le x < b)}$ **have the same value in CRV.**

Continuous Random Variables (CRV)

- The times taken (minutes) to share 5 GB sized files via a network are given in the table below.

Times taken (x)	f	f_i
62 – 63	460	0.114
63-64	750	0.186
64 – 65	970	0.241
65 – 66	760	0.189
66 – 67	640	0.159
67 - 68	440	0.109
	$\sum f = 4020$	$\Sigma f_i = 1.00$

Table 4.2.1: Frequency and relative frequency distributions of times taken to transfer 5 GB files.

Two characteristics of probability distribution for CRV:

1. Probability of x takes a value within an interval lies between 0 and 1.
$0 \leq \boldsymbol{P(a < x < b) \leq 1}$

2. The total probability of all the intervals within which x can take a value is 1.
$\boldsymbol{P(a < x < b) + P(a_1 < x < b_1) + \cdots + P(a_n < x < b_n) = 1}$

Continuous Random Variables (CRV)

- The times taken (minutes) to share 5 GB sized files via a network are given in the table below.

Times taken (x)	f	f_i
62 – 63	460	0.114
63-64	750	0.186
64 – 65	970	0.241
65 – 66	760	0.189
66 – 67	640	0.159
67 - 68	440	0.109
	$\sum f = 4020$	$\Sigma f_i = 1.00$

Table 4.2.1: Frequency and relative frequency distributions of times taken to transfer 5 GB files.

Example 01: Calculate the probability that x lies in the interval 65 to 67.

Solution: $P(65 < x < 68) = P(65 < x < 66) + P(66 < \mathrm{x} < 67)$

$= 0.189 + 0.159 = 0.348\ (Ans)$

Example 02: Calculate the probability that x = 64, and x = 66.

Solution: $\mathrm{P(x) = P(x{=}64) = P(x{=}66) = 0\ (Ans)}$

Continuous Random Variables (CRV)

- The times taken (minutes) to share 5 GB sized files via a network are given in the table below.

Times taken (x)	f	f_i
62 – 63	460	0.114
63-64	750	0.186
64 – 65	970	0.241
65 – 66	760	0.189
66 – 67	640	0.159
69 - 70	440	0.109
	$\sum f = 4020$	$\Sigma f_i = 1.00$

Table 4.2.1: Frequency and relative frequency distributions of times taken to transfer 5 GB files.

Example 03: Calculate the probability that a randomly selected value is within the range 65 to 67.

Solution: $P(65 < x < 68) = P(65 < x < 66) + P(66 < \mathrm{x} < 67)$

$= 0.189 + 0.159 = 0.348\ (Ans)$

Example 04: Calculate the probability that a randomly selected value is 64.

Solution: P(x) = P(x=64) = 0 (Ans)

PDF and CDF

- Look at the table at 4.2.1, there is a probability distribution (relative frequency distribution). When you plot the probability distribution of a continuous random variable, you get a curve.
- Every curve can be defined by a function.
- The function that is described by the plot of probability distribution from a continuous random variable is known as its Probability Density Function (PDF).
- When PDF is integrated with respect to its variable, we get the CDF (Cumulative Density Function).
- For discrete random variables, such function is known as Probability Mass Function (PMF), which is beyond the scope of this course.
- The probability density function (probability distribution curve) for Table 4.1.2 is drawn in the next page.

PDF and CDF

- Fig 4.1 illustrates the PDF for the continuous random variable X (time to send 5 GB files) as mentioned in table 4.2.1

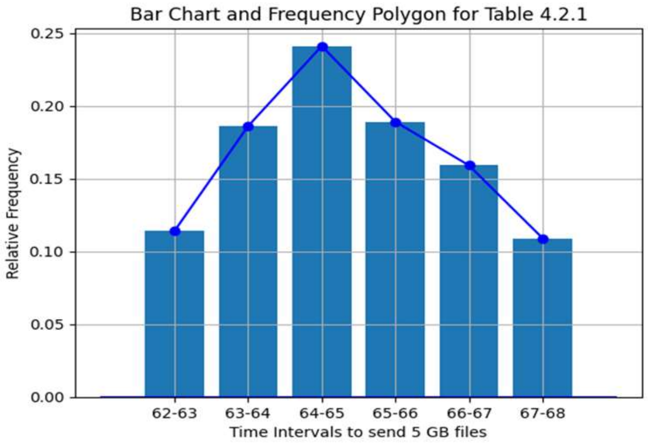


Fig 4.1: PDF of X shown using Bar Chart and Frequency Polygon

The frequency polygon is an approximation for the PDF. However, we could use Histogram instead of the Bar Chart if our data was large.

PDF and CDF

- Fig 4.2 illustrates the total area under the PDF for the continuous random variable X (time to send 5 GB files) as mentioned in table 4.2.1

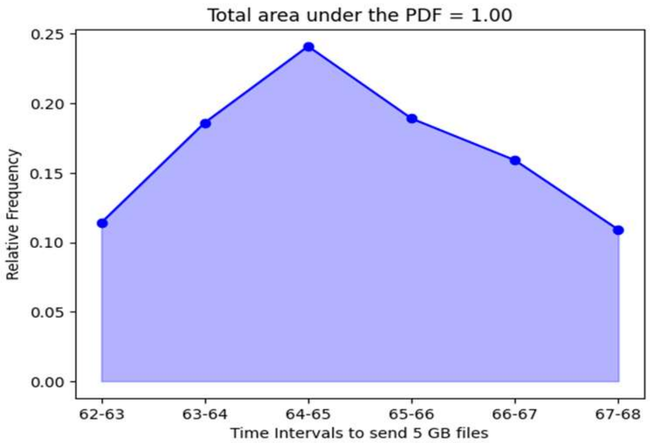


Fig 4.2: Total area under the probability distribution curve

The total area under the probability density curve is 1.00, because,

$$\boldsymbol{P(a < x < b) + P(a_1 < x < b_1) + \cdots + P(a_n < x < b_n) = 1}$$

In Fig 4.2, blue shaded area represents the total area under the probability density function.

PDF and CDF

- Fig 4.3 illustrates the probability of x takes on a random value between the interval a to b.

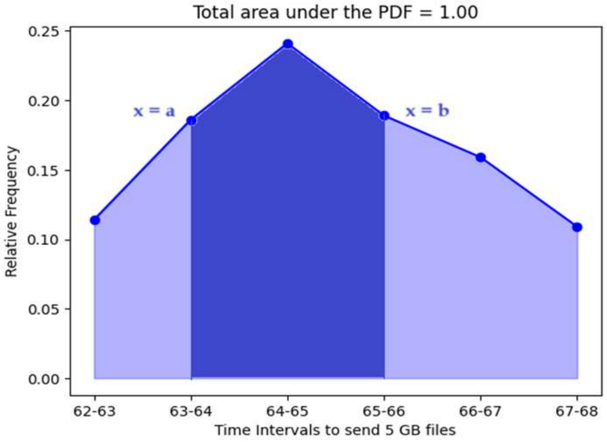


Fig 4.3: PDF of X shown using Bar Chart and Frequency Polygon

The **dark blue** shaded area under the curve is the probability of x takes a value -
Less than b and greater than a: $\boldsymbol{P(a < x < b)}$
Less than or equal to b and greater than or equal to a: $\boldsymbol{P(a \le x \le b)}$
Less than or equal to b and greater than a: $\boldsymbol{P(a < x \le b)}$
Less than b and greater than or equal to a: $\boldsymbol{P(a \le x < b)}$

Because in CRV, $\boldsymbol{P(a < x < b) = P(a \le x \le b) = P(a < x \le b) = P(a \le x < b)}$

PDF and CDF

- Fig 4.3 illustrates the probability of x takes on a random value between the interval $\mathbf{a}$ to $\mathbf{b}$.

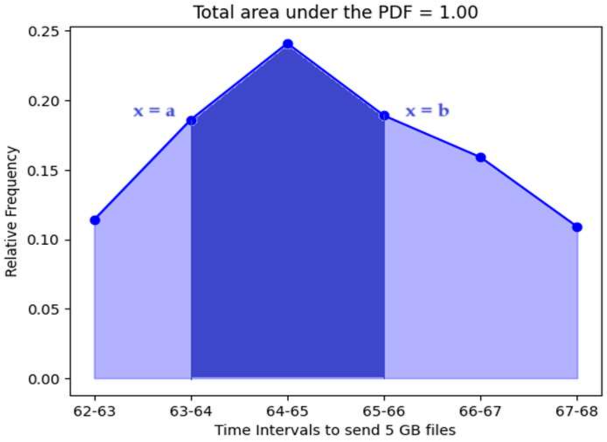


Fig 4.3: PDF of X shown using Bar Chart and Frequency Polygon

In Fig 4.3, given that, a = 63, and b = 65. Hence, from table 4.2.1,
$P(a < x < b) = P(63 < x < 65) = 0.186 + 0.241 = 0.427$

However, the probability that x takes on a single value is 0. Therefore,
$P(x = a) = P(x = b) = P(x = 63) = P(x = 65) = 0$

PDF and CDF

- For any continuous random variable **X**,

1. If you integrate PDF over the range (a,b), you get the probability of x taking a value within the range (a,b). $\int_a^b f(x)dx = P(a \leq x \leq b)$
2. If you integrate PDF over the range (0,x), you get the CDF. $F(x) = \int_0^x f(x)dx$
3. If you plug in a value for $x = a$ in the CDF, you get the probability of x taking a value less than a. $F(a) = P(x \leq a)$
4. If you plug in a value for $x = a$ and $x = b$ in the CDF, the probability of x taking a value within the range (a,b) is $F(b) - F(a)$. $P(a < x \leq b) = F(b) - F(a)$
5. Mean, $\mu = E(x) = \int_{-\infty}^{+\infty} xf(x)dx$
6. Variance, $\sigma^2 = E(x^2) - \left(E(x)\right)^2$
7. Standard Deviation, $\sigma = \sqrt{\sigma^2}$

BRAC UNIVERSITY

Inspiring Excellence

PDF and CDF

- **Example 01:** The PDF of a continuous random variable is $f(x) = \frac{1}{8}x.$

A. Find out CDF.

B. What is the probability that x takes on a value between 2 and 3?

Solution A: Given that, $f(x) = \frac{1}{8}x.$

So, $F(x) = \int \frac{1}{8}x\,dx = \frac{1}{8}.\frac{x^2}{2} = \frac{x^2}{16}$ (Ans)

Solution B: $P(2 < x < 3.5) = \int_2^3 \frac{1}{8}x\,dx = \left(\frac{3^2}{8}\right) - (\frac{2^2}{8}) = 0.3125\ \ (Ans)$

PDF and CDF

- **Example 02:** Suppose, X is a continuous random variable with $f(x) = \frac{1}{2}x$ for x in the interval [0,2] and $f(x) = 0$ elsewhere. Calculate E(x), and σ^2.

Solution: Given that, PDF: $f(x) = \begin{cases} \frac{1}{2}x \ if \ 0 \le x \le 2 \\ 0 \qquad elsewhere \end{cases}$

$$E(x) = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^2 x.\frac{1}{2}xdx = \int_0^2 \frac{x^2}{2}dx = \frac{2^3}{6} - \frac{0^2}{6} = \frac{4}{3} \text{ (Ans)}$$

$$E(x^2) = \int_{-\infty}^{+\infty} x^2 f(x)dx = \int_0^2 x^2.\frac{1}{2}xdx = \frac{2^4}{8} = 2$$

Now, $\sigma^2 = E(x^2) - \mathrm{E(x)}^2 = 2 - \left(\frac{4}{3}\right)^2 = 0.22$ (Ans)

PDF and CDF

- **Example 03:** The PDF of a continuous random variable X is given. Calculate CDF, expected value, and variance. PDF: $f(x) = \begin{cases} \frac{2x}{3} & if\ 0 \le x \le 1 \\ \frac{2}{3} & if\ 1 \le x \le 2 \\ 0 & elsewhere \end{cases}$ Also, calculate $P(0.25 < x < 0.95), and\ P(1.25 < x < 1.75)$.

Solution: We get CDF, by integrating PDF.

$$F(x) = \begin{cases} \frac{x^2}{3} & if\ 0 \le x \le 1 \\ \frac{2}{3}x & if\ 1 \le x \le 2 \\ 0 & elsewhere \end{cases}$$ (Ans)

PDF and CDF

- $E(x) = \int_{-\infty}^{+\infty} xf(x)dx = E(x) = \int_0^1 x.\frac{2x}{3}dx + \int_1^2 x.\frac{2}{3}dx = \frac{11}{9}(Ans)$

$E(x^2) = \int_{-\infty}^{+\infty} x^2f(x)dx = \int_0^1 x^2.\frac{2x}{3}dx + \int_1^2 x^2.\frac{2}{3}dx = \frac{31}{18}$

Variance = $\frac{31}{18} - \left(\frac{11}{9}\right)^2 = 0.228\ (Ans)$

To calculate $P(0.25 < x < 0.95)$, we would use the first CDF which is valid between 0 to 1. Because 0.25 and 0.95 falls under the range 0 to 1.

To calculate $P(1.25 < x < 1.75)$, we would use the second CDF which is valid between 1 to 2. Because 1.25 and 1.75 falls under the range 1 to 2.

- $P(0.25 < x < 0.95) = \frac{0.95^2}{3} - \frac{0.25^2}{3} = 0.28\ (Ans)$
- $P(1.25 < x < 1.75) = \frac{2}{3} * 1.75 - \frac{2}{3} * 1.25 = 0.33\ (Ans)$

BRAC UNIVERSITY

Inspiring Excellence

Continuous Probability Distributions

- Two types of continuous probability distributions:

1. Exponential Distribution
2. Normal Distribution

In this chapter, we will only learn about exponential distribution. We will learn about normal distribution in the next chapter.

Exponential Distribution

- The exponential distribution is a continuous probability distribution commonly used to model the time between events in a Poisson process, where events occur continuously and independently at a constant average rate.
- The PDF of an exponential distribution is given by, $\boldsymbol{f(x;\lambda) = \lambda e^{-\lambda x}\ for\ x \geq 0}$
- The CDF of an exponential distribution is given by, $\boldsymbol{F(x;\lambda) = 1 - e^{-\lambda x}\ for\ x \geq 0}$
- **The CDF is obtained** by integrating PDF. $\boldsymbol{F(x) = \int_0^x f(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}}$
- Conversely, **the PDF is obtained** by differentiating CDF.
- Mean, $\boldsymbol{\mu = \frac{1}{\lambda}}$
- Variance, $\boldsymbol{\sigma^2 = \frac{1}{\lambda^2}}$
- Standard Deviation, $\boldsymbol{\sigma = \sqrt{\sigma^2}}$

- For calculating $\boldsymbol{\mu}$, $\boldsymbol{\sigma^2}$, and $\boldsymbol{\sigma}$, you can use the formulas mentioned in Page 29 as well, but the process will be lengthy and difficult.

BRAC UNIVERSITY

Inspiring Excellence

Exponential Distribution

- Example of some continuous random variables that follow exponential distribution:

1. **Lifetime of a Light Bulb:** The time until a light bulb burns out, assuming a constant failure rate, is often modeled with an exponential distribution.
2. **Waiting Time for a Bus:** If buses arrive at a station randomly but with a constant average rate, the waiting time for the next bus can be modeled as an exponentially distributed variable.
3. **Duration of a Phone Call:** In a call center, the duration of a call might follow an exponential distribution if calls are cut off at random.
4. **Time Between Failures in a Computer System:** The time between hardware or software failures in a computer system, assuming constant failure rates, can be modeled with an exponential distribution.

Exponential Distribution

- **How to understand whether a continuous variable follows exponential distribution?**

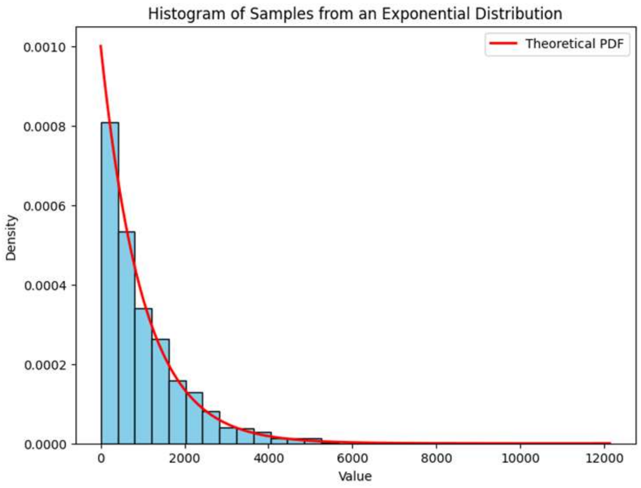


Fig 4.4: Identical histogram of exponential distribution

Exponentially distributed data are generally consisted of “time until an event” related data, as mentioned in the previous page.
When you’ll create histogram with such data, the histogram will look like Fig 4.4. It shows a right skewed distribution with a rapid decline from a peak at or near zero. Slide to next page for a detailed example.

Exponential Distribution

- Continuous data that represents “time until an event occurs”, usually follows exponential distribution. For example, battery lifetimes. The lifetimes of batteries follow an exponential distribution.
- Lets understand exponential distribution with an example experiment regarding quality control (QC). Suppose, we want to know the probability of a battery manufactured in XYZ industry fails between 700 to 900 hours. The threshold for passing the quality control test is that the probability needs to be less than 0.1. How can we do so?

1. First of all, we draw a representative sample of batteries from the industry.
2. We take the mean of lifetimes of these batteries. Assume, the mean lifetime is 1000 hours.
3. Make a histogram of **battery lifetimes** and you will see it follows exponential distribution.
4. Calculate the lambda parameter (λ) using the formula: $\lambda = \frac{1}{\mu}$.
5. Now, we have every data to write the PDF of exponential distribution. Use the PDF to calculate our desired probability. See example 1 to example 4 for details.

Exponential Distribution

- **Example 01:** Assume that, you are given a large sample on battery lifetimes. After making a histogram, you see that it follows an exponential distribution. Write the PDF for the "battery lifetimes" variable. The mean lifetime of the batteries from the sample is 1000 hours.

Solution: We know, $\mu = \frac{1}{\lambda}$. So, $\lambda = \frac{1}{1000} = 0.001$

Therefore, $f(x; 0.001) = 0.001e^{-0.001x}\ for\ x \geq 0$ (Ans)

Exponential Distribution

- **Example 02:** In the PDF, what do you get if you plug in x = 800 (hours)?

Solution: $f(800; 0.001) = 0.001e^{-0.001*800} = 0.00044$

The PDF value at x = 800 hours represents the likelihood density of the battery lasting precisely 800 hours. It is not the probability itself, but rather a density, which means that it gives an idea of how probable it is for the battery to last around 800 hours compared to other times.

BRAC UNIVERSITY

Inspiring Excellence

Exponential Distribution

- **Example 03:** How can you compute the probability that a randomly selected battery lasts between **700** to **900** hours?

Solution: By integrating the PDF with a lower bound of 0 and upper bound of x, we get the CDF.

However, right now, we skip this part because we already know that the CDF for a exponential distribution is, $F(x;\lambda) = 1 - e^{-\lambda x}\ for\ x \geq 0$

Therefore, CDF: $F(x; 0.001) = 1 - e^{-0.001x}$

By calculating F(900) – F(700), we can compute the probability.

However, we also can compute the probability by integrating PDF over the range 700 to 900.

Exponential Distribution

- **Example 04:** Calculate the probability that a randomly selected battery lasts between 700 to 900 hours? Can it pass the QC test is desired probability is less than 10%?

Solution: $P(700 < x < 900) = F(900) - F(700) = (1 - e^{-0.001*900}) - (1 - e^{-0.001*700})$

$= 0.593 - 0.503 = 0.09$ (Ans)

0.09 (9%) < 0.1(10%). Hence, the batteries passed the QC test. (Ans)

- **Example 05:** Calculate variance and SD.

Solution: Variance, $\sigma^2 = \frac{1}{0.001^2} = 1000000$

Hence, SD, $\sigma = \sqrt{1000000} = 1000$

- In exponential distributions, the mean and standard deviation are always the same.

STA201: Elements of Statistics and Probability
Department of MNS, BRAC University

Chapter 05: Normal Distribution and Hypothesis Testing

KAZI SAKIB HASAN (24341237)

COMPUTER SCIENCE PROGRAM, BRAC UNIVERSITY

kazi.sakib.hasan@g.bracu.ac.bd

Brief Contents

NORMAL DISTRIBUTION

STANDARD NORMAL DISTRIBUTION

- Z-SCORE
- DETECTING OUTLIERS

CALCULATING PROBABILITIES

- NORMAL PROBABILITY TABLE

HYPOTHESIS TESTING

- TYPES OF HYPOTHESIS TESTING
- HYPOTHESIS TESTING (ONE SAMPLE) WITH Z-TEST
  - P-VALUE APPROACH
  - CRITICAL POINT APPROACH

Normal Distribution

- In the previous chapter, we studied about exponential distribution, a type of continuous probability distribution. In this chapter, we will learn about the other continuous probability distribution, which is the normal distribution (Gaussian distribution).
- In the pre-requisites, we had the basic knowledge about normal distribution. In this chapter, our learning focus will be confined within:

1. Calculating probabilities using normal probability table.
2. Calculating z-scores.
3. Hypothesis testing using normal distribution.

Normal Distribution

- **Review:** When you draw the histogram for a data (e.g. heights of students in a classroom), and find out the histogram looks symmetrical, or the skewness is very close to 0, we call that the data is normally distributed (e.g. the heights of students is a normally distributed data). This normally distributed data is symmetrical in shape. and is defined by $X \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, where **X = dataset/variable name** (e.g. heights), $\boldsymbol{\mu}$ **= mean**, $\boldsymbol{\sigma}$ **= SD**.

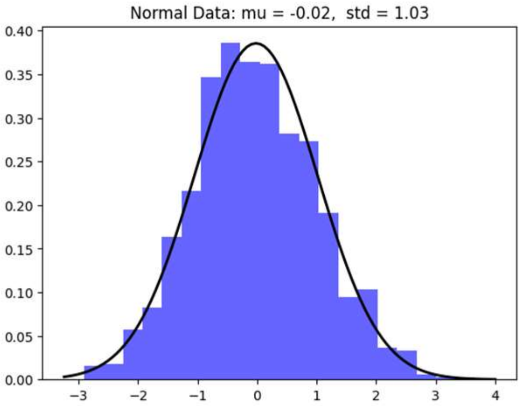


Fig 5.1: Shape of a normally distributed dataset

Fig 5.1 depicts a normally distributed univariate dataset with mean = -0.02 and standard deviation = 1.03. The shape of the data is symmetrical.

If a continuous variable is defined by $\boldsymbol{X}$, then we write, $X \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. It means, $\boldsymbol{X}$ is normally distributed with mean $\boldsymbol{\mu}$ and SD $\boldsymbol{\sigma}$.

In Fig 5.1, $\boldsymbol{X} \sim \boldsymbol{N}(-\mathbf{0.02}, \mathbf{1.03})$

Standard Normal Distribution

- If the mean and standard deviation of a normally distributed data is standardized with 0 and 1 respectively, then we call it **standard normal distribution**. Therefore, a standard normal distribution is defined as $\boldsymbol{X \sim N(0, 1).}$
- **Now, how do we standardize the normal distribution?**

- By calculating the z-score for individual data points and replace the z-scores with their corresponding data points.
- z-score = $\frac{x-\mu}{\sigma}$
- Also represented as, z-score = $\frac{observed\ value\ - mean}{standard\ deviation}$

BRAC UNIVERSITY
Inspiring Excellence

Standard Normal Distribution

- Assume, a continuous variable (or, dataset) $X = [x_1, x_2, x_3, \dots\dots, x_n]\,, where\ X \sim N(5, 2.5)$

If we calculate z-score for each of the data point:

$$z_{x_1} = \frac{x_1 - 5}{2.5}\ , z_{x_2} = \frac{x_2 - 5}{2.5}, z_{x_3} = \frac{x_3 - 5}{2.5}\ , \dots\dots\dots, z_{x_n} = \frac{x_n - 5}{2.5}$$

We will see that, the new dataset with these z-scores for their respective x_i is also normally distributed, where mean is 0 (or, very close to 0) and SD is 1.

Hence, $Z = [z_{x_1}, z_{x_2}, z_{x_3}, \dots\dots, z_{x_n}]\,, where\ Z \sim N(0, 1)$

If z_i is the corresponding z-score for a data point x_i,

Then, $\boldsymbol{P(x < x_i) = P(z < z_i)}$

So do, $\boldsymbol{P(x > x_i) = P(z > z_i)}$

Standard Normal Distribution

- **Example 01:** Synthia and Ashiq are from different sections. The mean score of STA201 midterm exam in Synthia's section is 15.25 with a SD of 3.85. In Ashiq's section, the mean score is 19.75 with a SD of 2.5. Ashiq scored 21 out of 25 in the exam, and Synthia scored 19 out of 25. Who performed better? The score in both section follow a normal distribution.

Solution: $z_{synthia} = \frac{19-15.25}{3.85} = 0.97$

$z_{ashiq} = \frac{21-19.75}{2.5} = 0.5$

So, while Synthia is 0.97 SD far away from the mean, Ashiq is 0.5 SD far away the mean. Therefore, Synthia performed better than Ashiq in the exam. In standard normal distribution, Synthia's score is 0.97, and Ashiq's score is 0.5.

Standard Normal Distribution

- **Detecting Outliers**

According to the empirical rule, 95% data points fall within 2 SD from the mean, and 99% data points fall within 3 SD from the mean. So, if the z-score of a data point in any normally distributed dataset is more than 2, the data point is an outlier in that dataset.

However, for certain experiments, the threshold for being an outlier can be 3 SD, instead of 2 SD.

Example 02: Assume a continuous variable, X~N(50,9.85). Is 82 an outlier in the dataset X? Given that, in this particular experiment, if the z-score of a value exceeds 3, then its an outlier.

Solution: $z_{82} = \frac{82-50}{0.25} = 3.24 > 3$

So, 82 is an outlier in the dataset X.

Calculating Probabilities

- We know that in a continuous probability distribution, if we want to calculate probabilities, then we need to integrate its PDF over the desired interval.
- The PDF for a normal distribution is, $\boldsymbol{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\right)\left(\frac{x-\mu}{\sigma}\right)^2}}$
- So, assume that a normally distributed data is defined by $\boldsymbol{X \sim N(6, 2.25)}$, and we have to find the probability that $\boldsymbol{x}$ takes on a value between 5 to 7, then we need to integrate the PDF over the range (5,7), which will be very pathetic to perform.
- Fortunately, there is a table available entitled as normal probability table, that will help us to calculate probabilities without integrating the PDF. The table is accessible from the Appendix section of this chapter.

Calculating Probabilities

- **Which probabilities do we calculate using the table?**

1. The probability of x takes on a value between the range (a,b). $\boldsymbol{P(a \le x \le b)}$
2. The probability of x takes on a value less than a. $\boldsymbol{P(x \le a)}$
3. The probability of x takes on a value more than a. $\boldsymbol{P(x \ge a)}$
4. The probability that x takes on a single value is always 0. $\boldsymbol{P(x = a) = 0}$

A necessary complementary rule:

$P(x > a) = 1 - P(x < a)$

Because you cannot calculate $\boldsymbol{x}$ takes on a value more than $\boldsymbol{a}$: $\boldsymbol{P(x \ge a)}$, using the normal table.

In the next section, we will learn to compute and interpret case 1: $\boldsymbol{P(a \le x \le b)}$. The rest other cases are included in this single one.

Calculating $P(a \leq x \leq b)$

- **Review:** Since normal distribution is a continuous probability distribution,

$\boldsymbol{P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b)}$

For normal distribution,

$\boldsymbol{P(a \leq x \leq b) = P(a < x \leq b) = P(a \leq x < b) = P(a < x < b) = P(x < b) - P(x < a)}$

Scenario: Let, X~N(85,2.25) where X = accuracy of kNN models in different AI projects.

- Calculate the probability of getting an accuracy between **83** to **87**.
- Calculate that $\boldsymbol{x}$ takes on a value in between $\mathbf{83}$ to $\mathbf{87}$.
- Calculate $\boldsymbol{P(83 < x < 87)}$

The three statements stated above indicate the same question. Lets solve it step-by-step.

Calculating $P(a \leq x \leq b)$

- Fig 5.2 visualizes the problem statement.

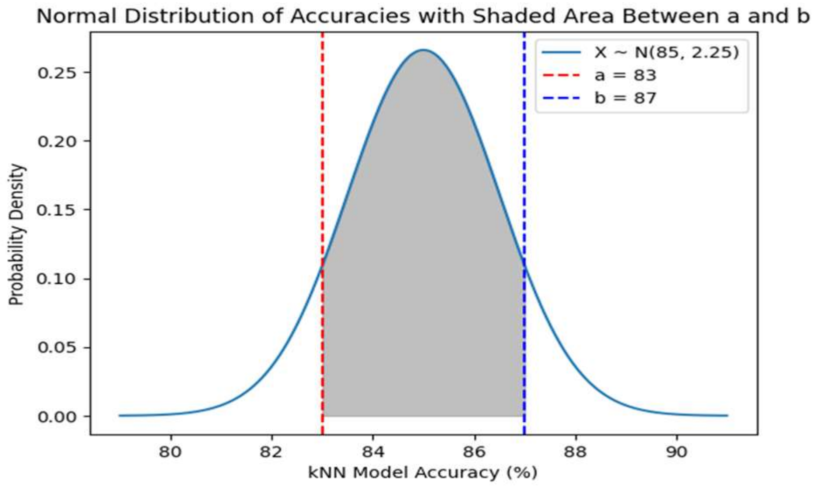


Fig 5.2: Shaded Area represents the value for $P(83 < x < 87)$.

Key points to notice in Fig 5.2:

$X \sim N(85, 2.25)$
Given, two observations, $(a, b) \ = \ (83,87)$

Entire area left to the red line a represents $P(x < 83)$
Entire area right to the red line a represents $P(x > 83)$

Entire area left to the blue line b represents $P(x < 87)$
Entire area right to the blue line b represents $P(x > 87)$

The **grey** shared area represents $P(83 < x < 87)$

BRAC UNIVERSITY
Inspiring Excellence

Calculating $P(a \leq x \leq b)$

- Fig 5.2 visualizes the problem statement.

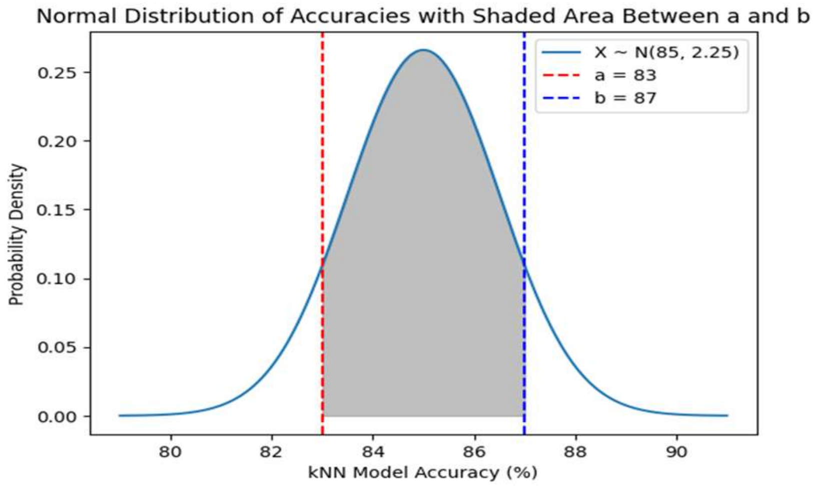


Fig 5.2: Shaded Area represents the value for $P(83 < x < 87)$.

We know, $P(a < x < b) = P(x < b) - P(x < a)$

Calculating $P(x < b)$

Step 01: z-score of b = $z_b = \frac{87-}{2.25} = 0.89$

Now, $P(x < b) = P(z_b < 0.89)$

Step 02: 0.89 = 0.8+0.09

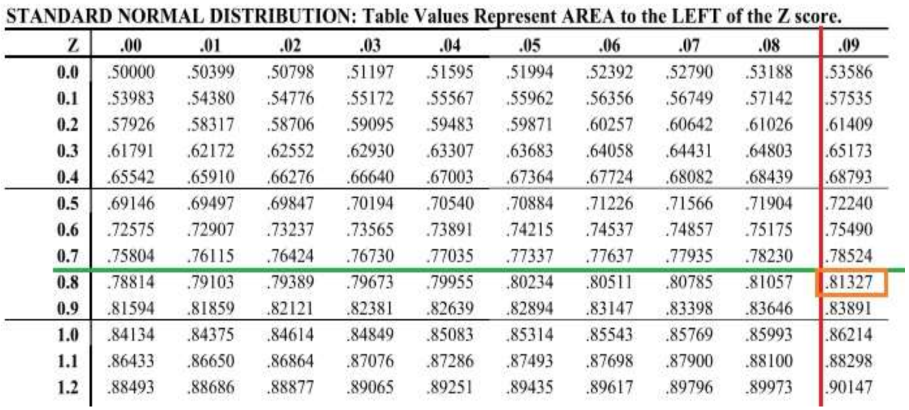
Look at the normal table attached in Appendix section,

- Find for 0.8 in the row titled as **z.**
- Find for 0.09 in the columns.

The intersecting point is the value for $P(x < b)$ = $P(z_b < 0.89)$, which is 0.81327.

Calculating $P(a \leq x \leq b)$

- Fig 5.3 shows how to calculate $P(z_b < 0.89)$



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147

Fig 5.3: Calculating $P(z_b < 0.89)$

We know, $P(a < x < b) = P(x < b) - P(x < a)$

Calculating $P(x < b)$

Step 01: z-score of b = $z_b = \frac{87-85}{2.25} = 0.89$

Now, $P(x < b) = P(z_b < 0.89)$

Step 02: $0.89 \; = \; 0.8 + 0.09$

Look at the normal table attached in the Appendix section,

- Find for 0.8 in the row titled as **z.**
- Find for 0.09 in the columns.

The intersecting point is the value for $P(x < b)$ = $P(z_b < 0.89)$, which is 0.81327.

Calculating $P(a \leq x \leq b)$

- Fig 5.2 visualizes the problem statement.

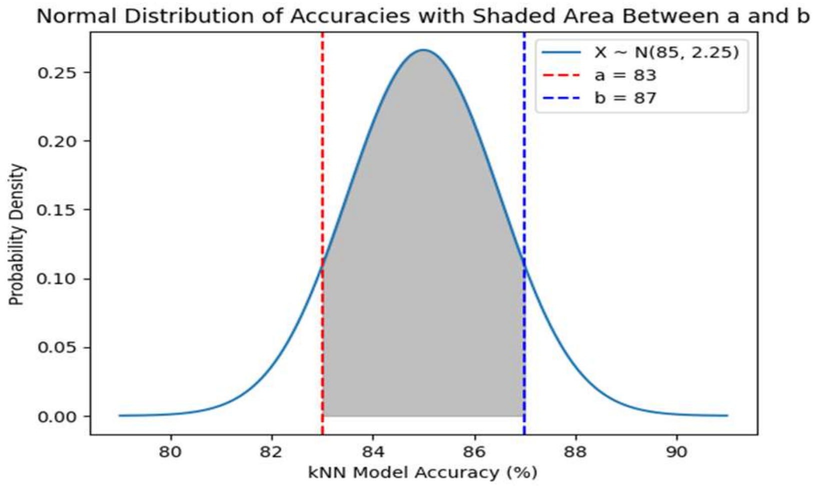


Fig 5.2: Shaded Area represents the value for $P(83 < x < 87)$.

We know, $P(a < x < b) = P(x < b) - P(x < a)$

Calculating $P(x < a)$

Step 01: z-score of a = $z_a = \frac{83-85}{2.25} = -0.89$

Now, $P(x < a) = P(z_a < -0.89)$

Step 02: -0.89 = - 0.8 - 0.09

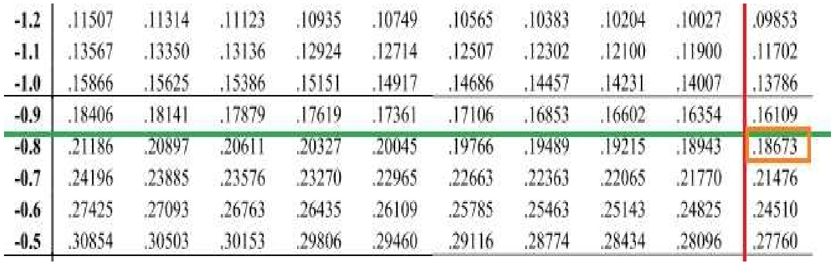
Look at the normal table attached in Appendix,

- Find for -0.8 in the row titled as **z.**
- Find for 0.09 in the columns. Don't care about the negative sign.

The intersecting point is the value for $P(x < a)$ = $P(z_b < -0.89)$, which is 0.18673.

Calculating $P(a \leq x \leq b)$

- Fig 5.4 shows how to calculate $P(z_a < -0.89)$



-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	$\boxed{.18673}$
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760

Fig 5.4: Calculating $P(z_a < -0.89)$

We know, $P(a < x < b) = P(x < b) - P(x < a)$

Calculating $P(x < a)$

Step 01: z-score of a = $z_a = \frac{83-85}{2.25} = -0.89$

Now, $P(x < a) = P(z_a < -0.89)$

Step 02: -0.89 = - 0.8 - 0.09

Look at the normal table attached in Appendix,

- Find for -0.8 in the row titled as **z.**
- Find for 0.09 in the columns. Don't care about the negative sign.

The intersecting point is the value for $P(x < a)$ = $P(z_b < -0.89)$, which is 0.18673.

Calculating $P(a \leq x \leq b)$

- Fig 5.2 visualizes the problem statement.

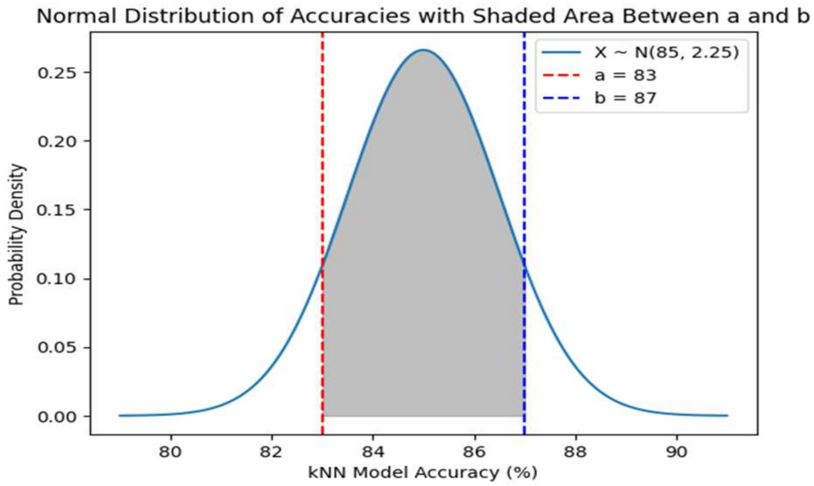


Fig 5.2: Shaded Area represents the value for $P(83 < x < 87)$.

We know, $P(a < x < b) = P(x < b) - P(x < a)$

We found out that,
$P(x < 87) = 0.8133$
$P(x < 83) = 0.1867$
So, $P(83 < x < 87) = 0.81327 - 0.18673 = 0.62654$
(Ans)

Calculating $P(a \leq x \leq b)$

- Fig 5.2 visualizes a scenario. Calculate $P(x > 87)$.

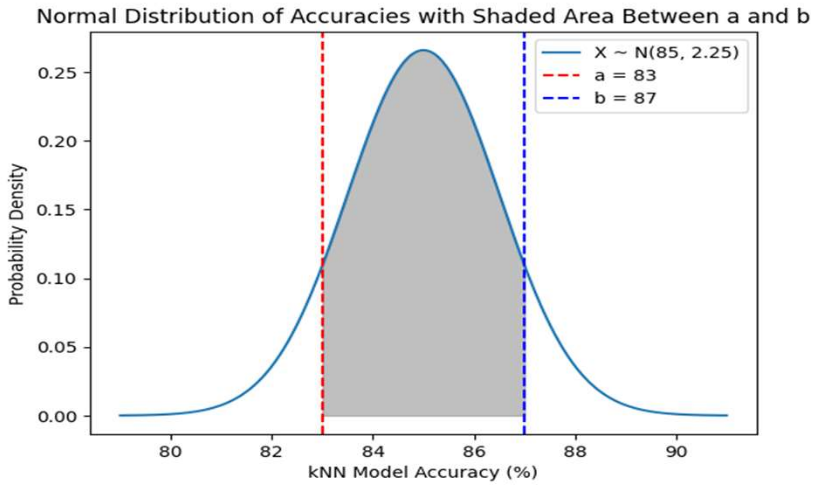


Fig 5.2: Shaded Area represents the value for $P(83 < x < 87)$.

We know, $P(x > b) = 1 - P(x < b)$

We found out that,
$P(x < 87) = 0.8133$
$P(x > 87) = 1 - P(x < 87) = 1 - 0.8133 = 0.1867$
(Ans)

Calculating $P(a \le x \le b)$

- Fig 5.5 visualizes a scenario. Find the cut off point for lowest 5% accuracies and interpret.

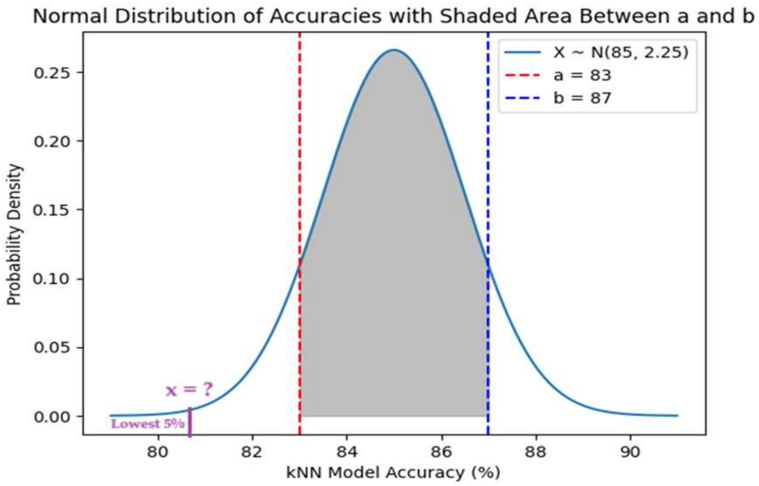


Fig 5.5: $\boldsymbol{x}$ is the cutoff point for lowest 5% accuracies.

Given, $P(z < z_i) = 0.05$

Now, $z_i = \frac{x_i - mean}{sd}$

So, $x_i = (z_i * sd) + mean$

$x_i = (z_i * 2.25) + 85$

The corresponding value of $z_i$for the probability 0.05 is -1.64 from the normal probability table.

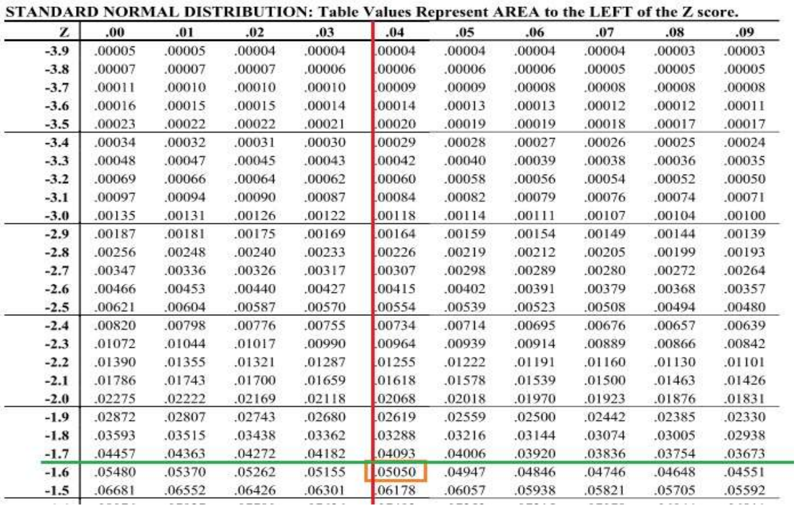
So, $x_i = (-1.64 * 2.25) + 85 = 81.31$

95% accuracies in the dataset is greater than or equal to 81.31% accuracy. Conversely, 5% accuracies are lower than this value.

Practical Interpretation: In 95% projects, the kNN model performed with 81.31% accuracy.

Calculating $P(a \le x \le b)$

- Fig 5.6 visualizes the approach to find the z-score for the probability value 0.05.



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592

Fig 5.6: Finding z-score for the probability value 0.05

Given, $P(z < z_i) = 0.05$

Now, $z_i = \frac{x_i - mea}{sd}$

So, $x_i = (z_i * sd) + mean$

$x_i = (z_i * 2.25) + 85$

The corresponding value of $z_i$for the probability 0.05 is -1.64 from the normal probability table.

So, $x_i = (-1.64 * 2.25) + 85 = 81.31$

95% accuracies in the dataset is greater than or equal to 81.31% accuracy. Conversely, 5% accuracies are lower than this value.

Practical Interpretation: In 95% projects, the kNN model performed with 81.31% accuracy.

Calculating $P(a \leq x \leq b)$

- **Example 01:** Assume that a data scientist of Facebook sampled 5000 accounts run by men of age 25-30 in Dhaka city and recorded the amount of time (minutes) they spend on stalking girls younger than them in a month. Let, X~N(25,2.75), where X = stalking time.

A. Interpret $X{\sim}N(25{,}2.75)$

B. Calculate $P(22 < x < 31)$ and interpret the findings.

C. Calculate $P(x > 27)$ and interpret.

Calculating $P(a \leq x \leq b)$

- **Solution A:** X is a normally distributed data with a sample size of 5000 that refers to the stalking time of 25-30 aged people. It follows a normal distribution with mean = 25 minutes and SD = 2.75 minutes.
- **Solution B:** $P(22 < x < 31) = P(x < 31) - P(x < 22)$

$z_{31} = \frac{31-25}{2.75} = 2.18\ (as,\ 2.1 + 0.08)$

From the normal probability table, $P(z < 2.18) = P(x < 28) = 0.9854$

$z_{22} = \frac{22-25}{2.75} = -1.09\ (as, -1.0 - 0.09)$

From the normal probability table, $P(z < -1.09) = P(x < 22) = 0.1379$

So, $P(22 < x < 31) = 0.9854 - 0.1379 = 0.8475\ (Ans)$

Interpretation: 84.75% men of age 25-30 in Dhaka City stalk girls younger than their age for 22 to 31 minutes in a month while using Facebook.

Calculating $P(a \leq x \leq b)$

- **Solution C:** $P(x > 27) = 1 - P(x < 27)$

$z_{27} = \frac{27-2}{2.75} = 0.72\ (as, 0.7 + 0.02)$

From the normal probability table, $P(z < 0.72) = P(x < 27) = 0.7642$

So, $P(x > 27) = 1 - 0.7642 = 0.2358\ (Ans)$

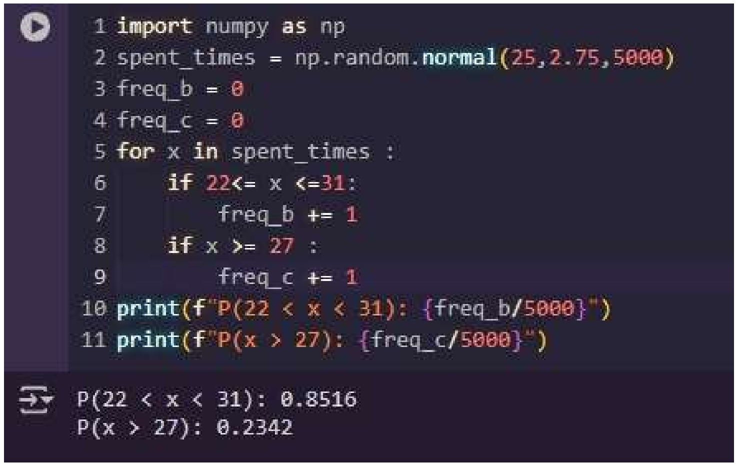
Interpretation: 23.58% men of age 25-30 in Dhaka City stalk girls younger than their age for more than or equal to 27 minutes in a month while using Facebook.

Calculating $P(a \leq x \leq b)$

- By the way, in real life, the data scientist of course would not use the table for such problems, because he already have the data. We were asked to calculate $P(22 < x < 31)$. Remember that, the sample size is 5000. So, if we count the frequency of data points that is within the range 22 to 31, and divide it by 5000, we will get the value for $P(22 < x < 31)$.
- The same approach is valid for the question C, where we were asked to calculate $P(x > 27)$. If we count the frequency of data points greater than or equal to 27, and divide it by 5000, we get $P(x > 27)$.
- In the next page, a synthetic normally distributed dataset is created using R and Python, and the value for $P(22 < x < 31)$ and $P(x > 27)$ is computed. Assume that, this synthetic dataset represents the data collected by the data scientist in Example 01.

Calculating $P(a \leq x \leq b)$

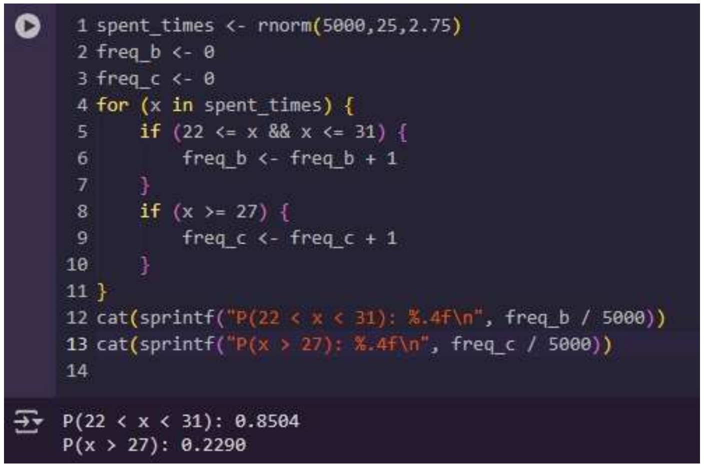
- Code snippet 5.1 and 5.2 shows that you can use the frequency count method to calculate probabilities, in a normally distributed dataset.



```python
import numpy as np
spent_times = np.random.normal(25,2.75,5000)
freq_b = 0
freq_c = 0
for x in spent_times :
    if 22<= x <=31:
        freq_b += 1
    if x >= 27 :
        freq_c += 1
print(f"P(22 < x < 31): {freq_b/5000}")
print(f"P(x > 27): {freq_c/5000}")
```

```
P(22 < x < 31): 0.8516
P(x > 27): 0.2342
```

Code snippet 5.1: Procedural approach to calculate normal probability in Python



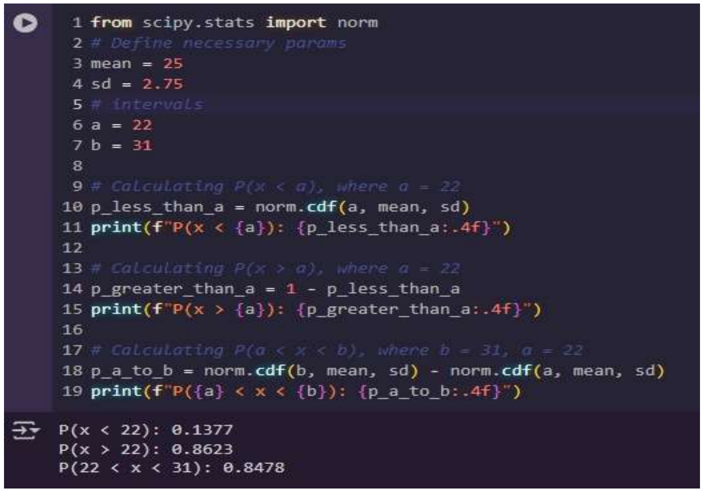
```r
spent_times <- rnorm(5000,25,2.75)
freq_b <- 0
freq_c <- 0
for (x in spent_times) {
    if (22 <= x && x <= 31) {
        freq_b <- freq_b + 1
    }
    if (x >= 27) {
        freq_c <- freq_c + 1
    }
}
cat(sprintf("P(22 < x < 31): %.4f\n", freq_b / 5000))
cat(sprintf("P(x > 27): %.4f\n", freq_c / 5000))
```

```
P(22 < x < 31): 0.8504
P(x > 27): 0.2290
```

Code snippet 5.1: Procedural approach to calculate normal probability in R

Calculating $P(a \leq x \leq b)$

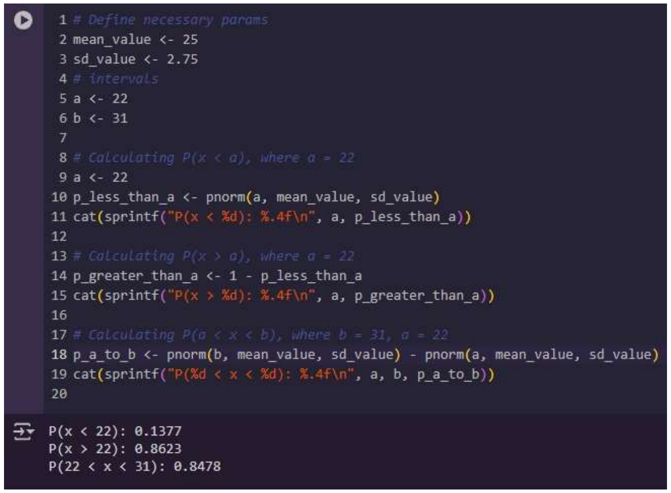
- However, this is not the exact way of calculating probability using programming software. An efficient version is attached below that would save your time in homework, or in personal researches.



```python
from scipy.stats import norm
# Define necessary params
mean = 25
sd = 2.75
# intervals
a = 22
b = 31

# Calculating P(x < a), where a = 22
p_less_than_a = norm.cdf(a, mean, sd)
print(f"P(x < {a}): {p_less_than_a:.4f}")

# Calculating P(x > a), where a = 22
p_greater_than_a = 1 - p_less_than_a
print(f"P(x > {a}): {p_greater_than_a:.4f}")

# Calculating P(a < x < b), where b = 31, a = 22
p_a_to_b = norm.cdf(b, mean, sd) - norm.cdf(a, mean, sd)
print(f"P({a} < x < {b}): {p_a_to_b:.4f}")
```

```
P(x < 22): 0.1377
P(x > 22): 0.8623
P(22 < x < 31): 0.8478
```

Code snippet 5.3: Efficient approach to calculate normal probability in Python



```r
# Define necessary params
mean_value <- 25
sd_value <- 2.75
# intervals
a <- 22
b <- 31

# Calculating P(x < a), where a = 22
a <- 22
p_less_than_a <- pnorm(a, mean_value, sd_value)
cat(sprintf("P(x < %d): %.4f\n", a, p_less_than_a))

# Calculating P(x > a), where a = 22
p_greater_than_a <- 1 - p_less_than_a
cat(sprintf("P(x > %d): %.4f\n", a, p_greater_than_a))

# Calculating P(a < x < b), where b = 31, a = 22
p_a_to_b <- pnorm(b, mean_value, sd_value) - pnorm(a, mean_value, sd_value)
cat(sprintf("P(%d < x < %d): %.4f\n", a, b, p_a_to_b))

```

```
P(x < 22): 0.1377
P(x > 22): 0.8623
P(22 < x < 31): 0.8478
```

Code snippet 5.4: Efficient approach to calculate normal probability in R

Inspiring Excellence

Calculating $P(a \leq x \leq b)$

- **Example 02:** The sample of K/D ratios of several COD Mobile players have been recorded. It is found that the K/D ratio follows a normal distribution with mean 2.23 and SD 1.36.

A. What is the cutoff point for the highest 3% K/D ratio? Interpret the result.

B. How many players have K/D ratio more than 3?

Calculating $P(a \leq x \leq b)$

- **Solution A:** Given, $P(z < z_i) = 1 - 0.03 = 0.97$

Now, $z_i = \frac{x_i - mean}{sd}$

So, $x_i = (z_i * sd) + mean$

$x_i = (z_i * 1.36) + 2.23$

The corresponding value of $z_i$for the probability 0.97 is 1.89 from the normal probability table.

So, $x_i = (1.89 * 1.36) + 2.23 = 4.80\ (Ans)$

Interpretation: The top 3% players with highest K/D ratio have greater than or equal to 4.80 K/D ratio. (Ans)

Calculating $P(a \leq x \leq b)$

- **Solution B:** $P(x > 3) = 1 - P(x < 3)$

$z_3 = \frac{3-2.23}{1.36} = 0.56$

$P(x < 3) = P(z < 0.56) = 0.7123$

$P(x > 3) = 1 - 0.7123 = 0.2877$

So, 28.77% of the players have more than 3 K/D ratio. (Ans)

Hypothesis Testing

- Hypothesis testing is a statistical analysis which is conducted to test the validity between two statements.
- One statement is considered to be **True**, which is the Null Hypothesis (H_0)
- The other statement is the contradictory statement to Null Hypothesis, which is the Alternative Hypothesis (H_1).
- In hypothesis testing, we always want to prove that Alternative Hypothesis is the valid one, whereas Null Hypothesis is the invalid statement.
- Statistically saying, we need enough evidence to reject the Null Hypothesis. Then automatically Alternative Hypothesis will be **True**.
- The p-value and the critical point describes whether we have enough evidences to reject the null hypothesis or not based on the comparison with level of significance.
- By saying "hypothesis testing" in this course, we are referring to **Hypothesis Testing about one Sample Mean.**

Types of Hypothesis Testing

- Two types of hypothesis testing :

1. Parametric tests
2. Non-parametric tests

We conduct parametric tests, when the underlying data is normally distributed. In this course, we assume that the data we are dealing with is normally distributed.

Types of Hypothesis Testing

- **Parametric tests:** We conduct it when the underlying data is normally distributed.

1. Hypothesis testing about one sample mean.

- z-test (when population SD is known)
- t-test (when population SD is not known)

2. Hypothesis testing about two sample mean.

- z-test (population SD is known and sample size > 30)
- t-test (population SD unknown)
  - Independent t-test
  - Matched pair t-test

3. Hypothesis testing about multiple sample mean.

- ANOVA

- In STA201, we will only study about Hypothesis testing about one sample mean that falls under the category of parametric tests.

Types of Hypothesis Testing

- There are two approaches to conduct hypothesis test for one sample mean:

1. The p-value approach.
2. The critical point approach.

Critical point approach is relatively easier than the p-value approach. You can use any of the approaches to test the hypothesis. We will learn about both approaches in this chapter.

Hypothesis Testing (One sample)

- **Hypothesis Testing about One Sample Mean**

Let us try to understand when we conduct hypothesis testing for one sample mean. Assume that, a sample from a population is taken and the sample mean is $\bar{x}_1$. Another sample from the same population is taken and this time the sample mean is $\bar{x}_2$. Now, what is the most accurate point estimate for the population? In such cases, we conduct hypothesis testing about one sample mean.

Let us try to understand with a scenario. Please slide to the next page.

Hypothesis Testing (One sample)

- **Scenario 1:** Imagine, a baker of a loaf shop claims that his loafs are 10.5 inches large. You take a sample (n=15) of loafs and found out that the mean of the loafs is 9.75 inches.

In this context,

- Null Hypothesis, $H_0: \mu = 10.5$ (The baker's claim is True)
- Alternative Hypothesis, $H_1: \mu < 10.5$ (The baker's claim is False)

The p-value will describe whether we should reject the null hypothesis or not. Rejecting the null hypothesis means, alternative hypothesis is true, which proves the baker wrong. His loafs are not 10.5 inches larger, those are actually smaller than this. He lied.

p-value

- The **p-value** is the probability of observing a test statistic as extreme as, or more extreme than, the one observed in your the data, assuming that the null hypothesis is true.
- In easier words, p-value (probability value) is the probability of if multiple samples are taken from the population, then each sample will be as extreme as the sample that was used to make a conclusion about the null hypothesis.
- Lets understand p-value from the context of scenario 1.

p-value

- Assume that, the p-value of this test comes out as 0.04. It means, if the baker's claim was true ($if\ H_0\ was\ true$), there is only a 4% chance of obtaining a sample mean as different from 10.5 inches as the one we observed (9.75 inches), or more extreme.
- So, practically we can infer that there is only 4% chance that the baker's claim was true. Therefore, if the significance level for the test was 4% or less, then the null hypothesis will be rejected and the bakery will be proven wrong.
- Whether we should reject the null hypothesis or not, it depends on the comparison between the significance level of the test. If the p-value becomes equal or smaller than the significance level, we reject the null hypothesis.

Significance Level

- Significance level is defined by α. It is the probability of we rejected the null hypothesis, whereas it was actually true. $\boldsymbol{\alpha = P(H_0\ is\ rejected\ |H_0\ is\ True)}$
- Every hypothesis test are conducted under different significance level. Whenever the p-value becomes equal or smaller than the significance level, we reject the null hypothesis. **For some cases**, we **don't reject** the null hypothesis if **significance level and p-value are equal.**
- Assume that, the significance level for **scenario 1** is 5% (0.05). In that case, we reject the null hypothesis because the p-value is 0.04, which is smaller than the significance level.

Significance Level

- When the p-value is equal or smaller than the significance level, we say –

1. The null hypothesis (H_0) is rejected.
2. There is a strong statistical evidence that the size of loafs of that shop is **significantly** less than 10.75 inches.
3. The alternative hypothesis (H_1) is true.

Significance Level

- **Example 01:** A microbiologist claims that the average bacterial growth rate is 200 colonies or more per day due to a new growth medium. A sample of 25 bacterial cultures grown in the medium is taken and the mean is 180 colonies. Hence, the hypothesis is tested, where the p-value resulted in 0.1. Verify the microbiologist's claim under 5% significance level. Interpret p-value.

Solution: H_0: $\mu \geq 200$

H_1: $\mu < 200$

Given, $\boldsymbol{\alpha}$ = 5% = 0.05

p-value = 0.1

p-value > $\boldsymbol{\alpha}$. So, null hypothesis cannot be rejected, and the alternative hypothesis is false. With 95% confidence and enough statistical evidences, we can say that the mean growth rate is 200 colonies.

p-value interpretation: There is 10% probability that the sample mean would be different from 200 colonies, if the sample mean was really 200 colonies.

Tails

- **Scenario 2:** A manufacturer claims that the average size of ceiling fan blades made by them is 56 cm. A sample of 25 fans is selected and it is found out that:

1. Case 01: The sample mean of those 25 fans is less than 56 cm. (Left-tailed test)
2. Case 02: The sample mean of those 25 fans is more than 56 cm. (Right-tailed test)
3. Case 03: The sample mean of those 25 fans is not 56 cm. (Two tailed test)

Case 01 and Case 02 are also known as one tailed test.

Z-test with p-value

- When the data is normally distributed, and population SD σ is known, we use the z-test.
- Necessary formulas for z-test.

1. Standard error, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

2. z-test statistic, $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$

3. For a **left-tailed** test, p-value = $P(Z < z)$

4. For a **right tailed** test, p-value = $P(Z > z)$

5. For a **two-tailed** test, p-value = $2 * P(Z > |z|)$

6. Based on the p-value, for any significance level greater than or equal to the p-value, we reject the null hypothesis.

7. X% significance level = (100-x)% confidence level. If the significance level is 5%, then the confidence level is (100-5)% = 95% (0.95). We also say, confidence level = $(1 - \alpha)$

Z-test with p-value

- **Example 01:** A manufacturer claims that the average lifespan of a new type of capacitor is 5,000 hours. A sample of 20 capacitors is taken and the mean lifespan of the capacitors is 4650 hours. The population standard deviation is 598. Test the hypothesis under 5% significance level to verify the claim of the manufacturer.

Solution: $H_0\text{: } \mu = 5000$

$H_1\text{: } \mu < 5000$ (This is a left tail test)

Standard error, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{598}{\sqrt{20}} = 133.71$

$z_{4650} = \frac{4650-500}{133.71} = -2.61$

From the normal probability table,

$P(z < -2.61) \ = \ 0.00453$

So, the p-value is 0.00453 which is < 0.05. Hence, we reject the null hypothesis.

There is sufficient statistical evidence to conclude that the mean lifetime of the capacitors is less than 5000 hours.

BRAC UNIVERSITY

Inspiring Excellence

Z-test with p-value

- **Example 02:** A software developer team claims that the average execution time of their new algorithm is 100 milliseconds or less than this. A senior developer sampled 15 execution times and found out that the average runtime is 103 milliseconds. Assume that, the population SD of the runtimes is 2.25 milliseconds. Verify the claim of the software developers under 3% significance level.
- **Solution:** H_0: $\mu \leq 100$ milliseconds

H_1: $\mu > 100$ milliseconds (This is a right tail test)

Standard error, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.25}{\sqrt{15}} = 0.58$

$z_{4650} = \frac{103-100}{0.58} = 5.17$

From the normal probability table,

$P(z < 5.17) \ = \ 0.997$, so, p-value = $P(z > 5.17) = 1 - 0.997 = 0.003 < 0.03$

So, the p-value is 0.003 which is < 0.03. Hence, we reject the null hypothesis. The mean runtimes is

Significantly greater than 100 milliseconds.

BRAC UNIVERSITY

Inspiring Excellence

Z-test with p-value

- **Example 03:** A new circuit design is claimed to have an average power consumption of 12 watts. A sample of 15 circuits are drawn and the calculated power consumption is 13.75 watts. Now, the average power is required 12 watts only, not more or less than it. Test the hypothesis, whether the circuit provides 12 watts or more or less than it, under 5% significant level. Assume the SD of the population of circuit is 2.15.
- **Solution:** H_0: $\mu = 12$ watts

H_1: $\mu \neq 12$ watts (This is a two-tailed test)

Standard error, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.15}{\sqrt{15}} = 0.55$

$z_{4650} = \frac{13.75-12}{0.55} = 3.18$

From the normal probability table,

$P(z < 3.18) = 0.99926$, so $P(z > 5.17) = 1 - 0.99926 = 0.00074$.

So, p-value = $2(0.00074) = 0.00148 < 0.05$. So, we reject the null hypothesis. It means that, that the average power consumption is significantly different from 12 watts.

Critical Point

- Critical point defines the rejection region and non-rejection region in a hypothesis test. If the test statistic falls in the rejection region, then we reject the null hypothesis. Else, we fail to reject the null hypothesis.
- If, $|test\ statistic| \geq |critical\ point|$, reject the Null Hypothesis.

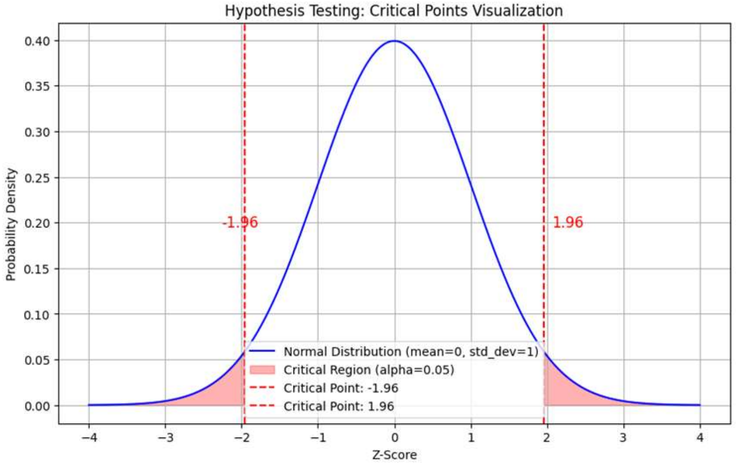


Fig 5.7: Critical points for a two-tailed test

Critical Point

- **How to find the critical point?**

For, **one** tailed hypothesis testing, the respective z-score for the predetermined confidence level (converted to probability value) in the normal probability table is the critical point. For example,

- What is the critical point for the significance level 5% $(\alpha = 0.05)$? (For one tailed hypothesis testing).

Confidence level = $(1 - \alpha) = (1 - 0.05) = 0.95$

In normal probability table, we can see the corresponding z-score for probability 0.95 is 1.64. Hence, the critical point for a 95% significance level in z-test is 1.64.

Critical Point

- **How to find the critical point?**

For, **two** tailed hypothesis testing, the respective z-score for the half of the predetermined confidence level (converted to probability value) in the normal probability table is the critical point. For example,

- What is the critical point for the significance level 5% $(\alpha = 0.05)$? (For two tailed hypothesis testing).

Confidence level = $(1 - \alpha/2) = (1 - 0.05/2) = 0.975$

In normal probability table, we can see the corresponding z-score for probability 0.975 is 1.96. Hence, the critical point for a 95% significance level in z-test is 1.96.

Z-test with Critical Point

- **Example 01:** A manufacturer claims that the average lifespan of a new type of capacitor is 5,000 hours. A sample of 20 capacitors is taken and the mean lifespan of the capacitors is 4650 hours. The population standard deviation is 598. Test the hypothesis under 5% significance level to verify the claim of the manufacturer.

Solution: H_0: $\mu = 5000$

H_1: $\mu < 5000$ (This is a left tail test)

Given, $\alpha = 5\% = 0.05$

$(1 - 0.05) = 0.95$

Standard error, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{598}{\sqrt{20}} = 133.71$

$z_{4650} = \frac{4650-500}{133.71} = -2.61$

From the normal probability table, z-score for 0.95 is 1.64. So, critical point is 1.64.

| -2.61| >| 1.64| . Hence, the z-test statistic is inside the rejection region. Therefore, we reject the null hypothesis.

There is sufficient statistical evidence to conclude that the mean lifetime of the capacitors is less than 5000 hours.

BRAC UNIVERSITY

Inspiring Excellence

Z-test with Critical Point

- **Example 02:** A software developer team claims that the average execution time of their new algorithm is 100 milliseconds or less than this. A senior developer sampled 15 execution times and found out that the average runtime is 103 milliseconds. Assume that, the population SD of the runtimes is 2.25 milliseconds. Verify the claim of the software developers under 3% significance level.
- **Solution:** H_0: $\mu \leq 100$ milliseconds

H_1: $\mu > 100$ milliseconds (This is a right tail test)

Given, $\alpha = 3\% = 0.03$

$(1 - 0.03) = 0.97$

Standard error, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.25}{\sqrt{15}} = 0.58$

$z_{4650} = \frac{103-100}{0.58} = 5.17$

From the normal probability table, the z-score for 0.97 is 1.89. This is the critical point.

5.17 > 1.89. Since the z-test statistic falls in the rejected region, we reject the null hypothesis.

The mean runtimes is significantly greater than 100 milliseconds.

Z-test with Critical Point

- **Example 03:** A new circuit design is claimed to have an average power consumption of 12 watts. A sample of 15 circuits are drawn and the calculated power consumption is 13.75 watts. Now, the average power is required 12 watts only, not more or less than it. Test the hypothesis, whether the circuit provides 12 watts or more or less than it, under 5% significant level. Assume the SD of the population of circuit is 2.15.
- **Solution:** H_0: $\mu = 12$ watts

H_1: $\mu \neq 12$ watts (This is a two-tailed test)

Given, $\alpha = 5\% = 0.05$

$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

$(1 - 0.025) = 0.975$

Standard error, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.15}{\sqrt{15}} = 0.55$

$z_{4650} = \frac{13.75-12}{0.55} = 3.18$

From the normal probability table, the critical point for 0.975 is 1.96.

Z-test statistic 3.18 > critical value 1.96. So, we reject the null hypothesis. It means that, that the average power consumption is significantly different from 12 watts.

Appendix

- Notes on normal distribution for a quick overview: https://docs.google.com/document/d/1aVE6tweCedXUJcNTU4iBsTcPb2aHf6WzLDHxzcIMi3c/edit?usp=sharing
- Notes on hypothesis testing for a quick overview:

STA201: Hypothesis Test without using p-value and normal table - Google Docs

- Normal Probability Table : https://drive.google.com/file/d/1urTXC3puiSXFsKQI5oszRXsxY4GhoCId/view?usp=sharing