DATA SCIENCE

GRADE X

Version 1.0

Microsoft

DATA SCIENCE

GRADE X

Student Handbook

ACKNOWLEDGMENT

Patrons

- Sh. Ramesh Pokhriyal 'Nishank', Minister of Human Resource Development, Government of India
- Sh. Dhotre Sanjay Shamrao, Minister of State for Human Resource Development, Government of India
- Ms. Anita Karwal, IAS, Secretary, Department of School Education and Literacy, Ministry Human Resource Development, Government of India Advisory

Editorial and Creative Inputs

- Mr. Manuj Ahuja, IAS, Chairperson, Central Board of Secondary Education

Guidance and Support

- Dr. Biswajit Saha, Director (Skill Education & Training), Central Board of Secondary Education
- Dr. Joseph Emmanuel, Director (Academics), Central Board of Secondary Education
- Sh. Navtez Bal, Executive Director, Public Sector, Microsoft Corporation India Pvt. Ltd.
- Sh. Omjiwan Gupta, Director Education, Microsoft Corporation India Pvt. Ltd
- Dr. Vinnie Jauhari, Director Education Advocacy, Microsoft Corporation India Pvt. Ltd.
- Ms. Navdeep Kaur Kular, Education Program Manager, Allegis Services India

Value adder, Curator and Co-Ordinator

- Sh. Ravinder Pal Singh, Joint Secretary, Department of Skill Education, Central Board of Secondary Education

ABOUT THE HANDBOOK

In today's world, we have a surplus of data, and the demand for learning data science has never been greater. The students need to be provided a solid foundation on data science and technology for them to be industry ready.

The objective of this curriculum is to lay the foundation for Data Science, understanding how data is collected, analyzed and, how it can be used in solving problems and making decisions. It will also cover ethical issues with data including data governance and builds foundation for AI based applications of data science.

Therefore, CBSE is introducing 'Data Science' as a skill module of 12 hours duration in class VIII and as a skill subject in classes IX-XII.

CBSE acknowledges the initiative by Microsoft India in developing this data science handbook for class X students. This handbook introduces the concept of distributions, identifying patterns, data merging with practical examples. The course covers the theoretical concepts of data science followed by practical examples to develop critical thinking capabilities among students.

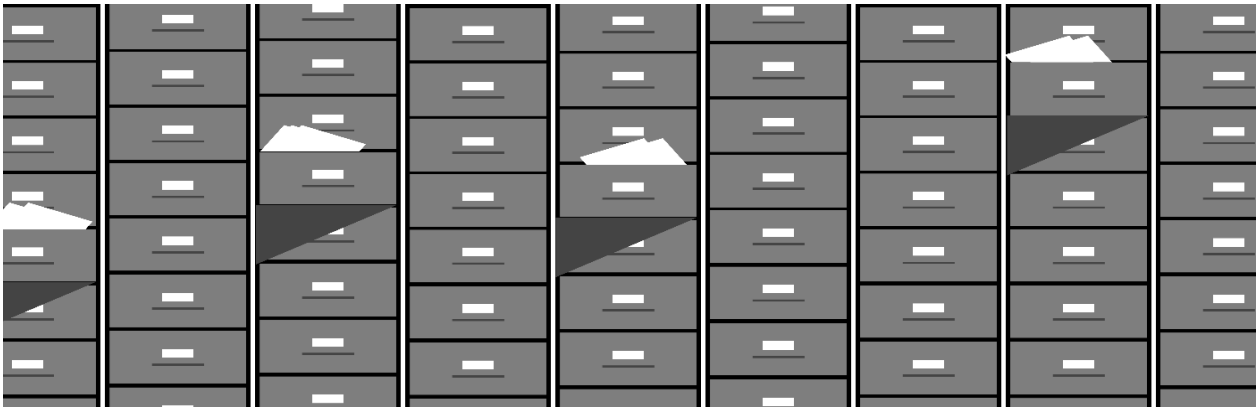
The purpose of the book is to enable the future workforce to acquire data science skills early in their educational phase and build a solid foundation to be industry-ready

Contents

CHAPTER

1

USE OF STATISTICS IN DATA SCIENCE



Studying this chapter should enable you to understand:

- What are subsets and relative frequency?
- Meaning of mean
- What is median and its usage in data science?
- What is mean absolute deviation?
- What is Standard Deviation?

1. Introduction

In the previous classes of data science, we have seen how data plays a vital role in our daily lives. We have also seen the significance of analyzing and visualizing the data. Now it is time to get into little more details of data analysis techniques and understand some of the statistical terminologies that are frequently used in data science. In this chapter, we will get to know some of the statistical concepts like what are subsets, what is mean, median and relative frequency? We will also see how these are used in the context of data science.

2. What are subsets?

Many a time, we encounter situations where we have a lot of data with us. However, for analysis, we do not need entire data for consideration.

Thus, instead of working with the whole data set, we can take a certain part of the data for our analysis. This division of a small set of data from a large set of data is known as a Subset.

Subsetting the data is a useful indexing feature for accessing object elements. It can be used for selecting and filtering variables and observations. We subset the data from a data frame to retrieve a part of the data that we need for a specific purpose. This helps us to

observe just the required set of data by filtering out unnecessary content.

For example, if you have a Table of 100 rows and 100 columns and you want to perform certain actions on the first 5 rows and the first 5 columns, you can separate it from the main table. This small table of 5 rows and 5 columns is known as a "Subset" in Data Analytics.

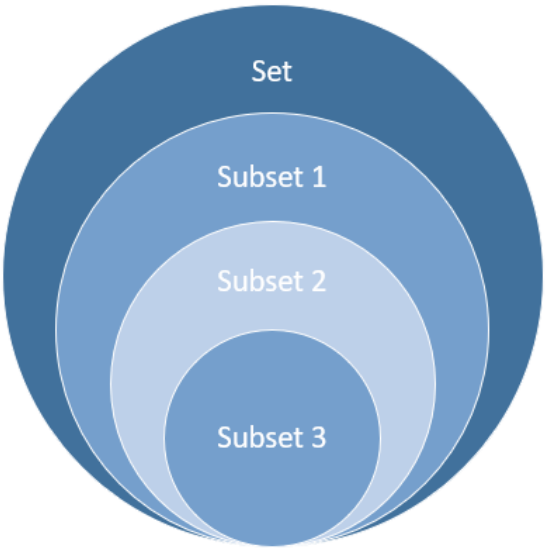


Fig 1.1 *Subset*

How do we subset the data?

Subsetting is a very significant component of data management and there are several ways that one can subset data. Let us now understand different ways of subsetting the data.

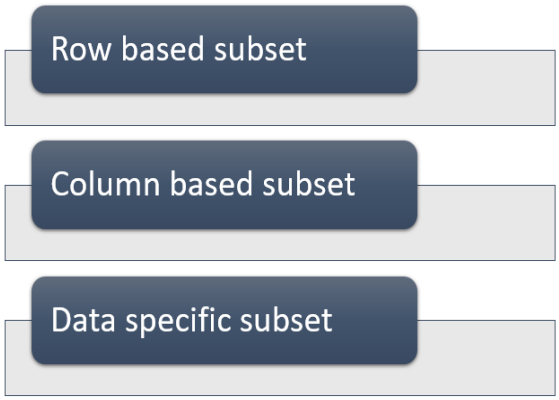


Fig 1.2 *Different ways to subset data*

Let us now try to understand each of them in more details

1. Row-based subsetting

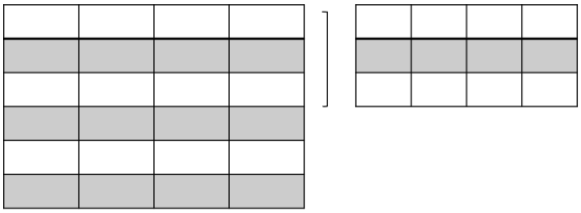


Fig 1.3 *Row based sub setting*

In this method of subsetting, we take some rows from the top or bottom of the table. Consider you have a table of 6 rows and 4 columns. You take the top 3 rows from the table.

2. Column based subsetting

Sometimes the original data set may contain a large number of columns and all of them may not be necessary to perform the analysis. We then select specific columns from the dataset. This process of subsetting is known as column-based subsetting.

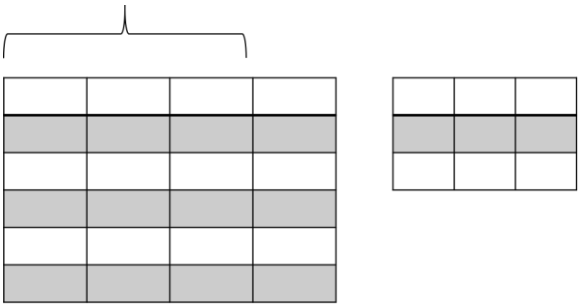


Fig 1.4 Column based sub setting

3. Data based subsetting

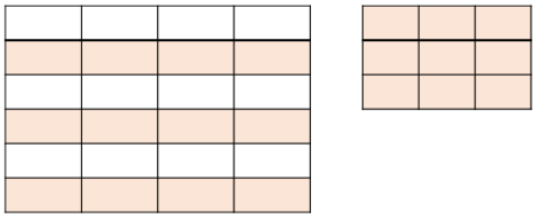


Fig 1.5 Data based sub setting

To subset the data based on specific data we use data-based subsetting. In the above figure, we have selected only those rows which are colored.

3.Two-way frequency table

Consider you are conducting a poll asking people if they like chocolates. You record the data in the below format.

Person name	Age	Like chocolates?
Person 1	6	Yes
Person 2	8	Yes
Person 3	13	Yes
Person 4	12	Yes
Person 5	18	No
Person 6	9	No
Person 7	16	Yes
Person 8	19	No
Person 9	14	No
Person 10	12	Yes

Fig 1.6 Poll on chocolates

If you now break down the data into age categories of (5 – 10 years), (10 – 15 years), and (15 – 20 years), and plot the number of people who liked and disliked chocolates then the table would look like below.

Age group	Like chocolates	Do not like chocolates
5 - 10	2	1
10 - 15	3	1
15 - 20	1	2

Fig 1.7 Two way frequency table

This type of table is called a two-way frequency table.

A two-way table is a statistical table that demonstrates the observed number or frequency for two variables, the rows indicate one category and the columns

indicate the other category. Two-way frequency tables show how many data points fit in each category. The row category in this example is "5-10 years", "10-15 years" and "15-20 years". The column category is their choice "Like chocolates" or "Do not like chocolates". Each cell tells us the number (or frequency) of the people.

There is a lot of information that we can get from this small table. For example,

How many people were questioned?
Answer: 10

How many people like chocolates?

Answer: 6

In which age group do people like chocolate the most?

Answer: 10 – 15

Let us now have a look at another example:

Example:

A survey of eighty people (40 men and 40 women) was taken on what genre of movie they would choose to watch, and the following responses were recorded:

- 8 men preferred comedy movies.
- 12 men preferred action movies.
- 14 men preferred horror movies.
- 16 women preferred comedy movies.
- 12 women preferred action movies.
- 6 women preferred horror movies.

The information collected is used to build the following two-way table:

Category	Comedy	Action	Horror	Total
Men	8	18	14	40
Women	23	10	7	40
Total	31	28	21	80

Fig 1.7(a) Two-way Table

4.Interpreting two-way tables

The entries in the table are counts. The table has several features:

- Categories are in the left column and top row
- The counts are placed in the center of the table.
- The totals are at the end of each row and column.
- A sum of all counts (a total) is placed at the bottom right

Example:

Category	Owns a car	Don't own a car	Total
Men	35	25	60
Women	40	20	60
Total	75	45	120

Fig 1.7(b) Two-way Table

In the above example, the rows of the table tell us whether it's a Male or a Female and the columns of the table tell

us if they own a car or not. Each cell tells us the number (or frequency) of people.

For example, 40 females own a car.

> ***Activity 1.1***
>
> Record how many of your friends like cricket and how many like football. Create a two-way relative frequency table with the data.

5.Two-way relative frequency table

Two-way relative frequency table very similar to the two-way frequency type of table. The only difference here is we consider percentage instead of numbers.

Two-way relative frequency tables represent what is the percentage of data points that fit in each category. We can take the help of row relative frequencies or column relative frequencies; it depends on the context of the problem.

Let us consider the below two-way table recording preferences of boys and girls

Preference	Girls	Boys
Indoor games	70	20
Outdoor games	30	80
Total	100	100

Fig 1.8 *Two way frequency table*

with regards to indoor and outdoor sports.

To convert this into a relative two-way frequency table we will convert individual cells into percentages.

Preference	Girls	Boys
Indoor games	70%	20%
Outdoor games	30%	80%
Total	100%	100%

Fig 1.9 *Two way relative frequency table*

Two-way relative frequency tables are helpful when there are different sample sizes in a dataset. Percentages make it easier to compare the preferences.

City	Temperature
Mumbai	21° C
Delhi	13° C
Chennai	24° C
Kolkata	15° C
Bangalore	20° C

Fig 1.10 *Calculating mean*

6.Meaning of mean

Mean is a measure of central tendency. In data science, Mean, also termed as the simple average, is an average value of a data set. Mean is a value in the data set around which the entire data is spread out. While mean is calculated, all values used in calculating the average are weighted equally.

The mean of a data set is calculated by adding up all the values in the data set and later dividing them by the number of values present in the data frame. Let us understand how to find a mean with the help of the below example.

Consider that we have a set of 11 numbers 10 to 20 in a data set.

Array = {10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}

So mean is calculated by adding up 10 numbers in the data set.

Sum of all the numbers = 165

Mean = 165/10 = 16.5

Let us now try to understand the real-life application of mean. We will create a data set for five cities in India for their minimum day temperature on a particular day. Let us record the temperatures in a table. Following is the sample of data that is collected:

To calculate the mean we will add the temperatures of all cities and divide it by 5.

Mean = (21 + 13 + 24 + 15 + 20)/5 = 18.5° C

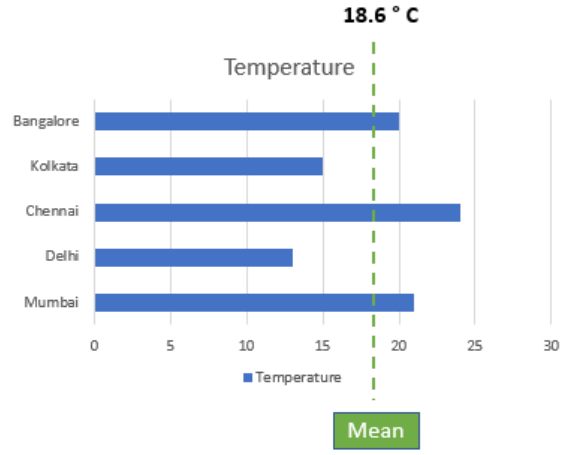


***Fig 1.11** Mean in a graph*

Activity 1.2

- Height of Ravi: 156cm
- Height of Juhi: 148cm
- Height of Shweta: 151cm
- Height of Kishan: 158cm

What is the mean?

Mean as represented in the above graph shows it stands for central tendency, meaning, it points to the center of the data set.

7.Median

The median like the mean is another form of central tendency. It is the middle point of a sorted data set.

To calculate the median, we must order our data set in ascending or descending order. If the data set is sorted from smallest value to biggest value, the exact middle value of the set is the Median.

Consider the below data set of 5 values.

Array = [12, 34, 56, 89, 32]

Now let us sort the data set.

Sorted array = [12, 32, 34, 56, 89]

The value at 3rd position is the middle point of the sorted list. So, 34 is our median for the array.

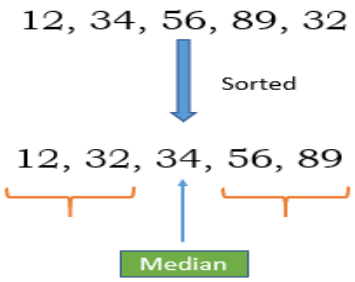


Fig 1.12 *Median*

In the previous example, we had a data set with an odd number of records. So, we could easily find out the middle point. But what if the data set has an even number of records? For these situations, there will be two middle points. Thus, we need to calculate the average of the two to get the median. The below example illustrates how to calculate median from an even number of records.

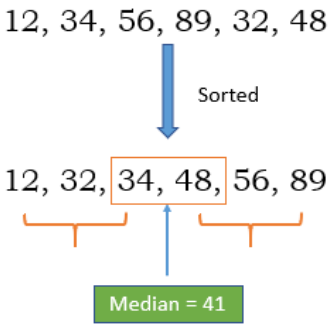


Fig 1.13 *Median for even number of records*

Mean vs median

So mean and median both represent the central tendency of a data set. So when do we use median over mean?

Median is a more accurate form of central tendency especially in scenarios where there are some irregular values also known as outliers. For example, consider the below scenario.

Your father gets his blood pressure checked every week. But due to some error in the device, the recording for one week was too high.

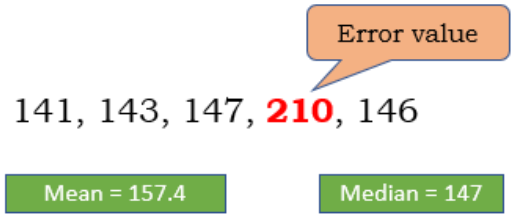


Fig 1.14 *Mean vs median*

So, for the above scenario we see the mean value deviates greatly from the regular blood pressure values due to a device error. Whereas the median value still accurately represents the central point of the data set. So under circumstances where there are outliers in the data set, median is a more effective measure of central tendency.

8.Mean Absolute Deviation

Mean **A**bsolute **D**eviation (MAD) is the average of how far away all values in a data set are from the mean.

Let understand this with an example. Consider the below data set.

12, 16, 10, 18, 11, 19

Step 1: Calculate the mean

Mean = (12 + 16 + 10 + 18 + 11 + 19) / 6 = **14** (rounded off)

Step 2: Calculate the distance of each data point from the mean. We need to find the absolute value. For example if the distance is -2, then we ignore the negative sign.

Below is the table after calculating the distance of each data point from the mean.

Value	Distance from mean (14)
12	2
16	2
10	4
18	4
11	3
19	5
Total	**20**

Fig 1.15 *Distance from mean*

|-2| = 2

Step 3:

Calculate the mean of the distances.

Mean of distances = (2 + 2 + 4 + 4 + 3 + 5) / 6 = **3.33**

So **3.33** is our mean absolute deviation, and the mean is **14**.

The value of Mean absolute deviation gives a very good understanding of the variability of the data set or in other words how scattered the data set is?

Activity 1.3

Calculate the mean absolute deviation for the below data set.

26, 35, 22, 28, 40, 38, 19

9. What is Standard Deviation?

The Standard Deviation is the measure of how spread out the numbers are. To be specific, standard deviation represents how much the data is spread out around the mean or an average.

For example, are all the points close to the average? Or are there lots of points way above or below the average?

In order to find standard deviation:

1. Calculate the mean by adding up all the data pieces and dividing it by the number of pieces of the data.
2. Subtract mean from every value
3. Square each of the differences

4. Find the average of squared numbers calculated in point number 3 to find the variance.
5. Lastly, find the square root of variance. That is the standard deviation.

For example, Take the values 1,2,3,5 and 8

Step 1: Calculate the mean

1+2+3+5+8 = 19

19/5 = 3.8 (mean)

Step 2: Subtract mean from every value

1- 3.8= -2.8

2- 3.8= -1.8

3- 3.8= -0.8

5- 3.8= 1.2

8- 3.8= 4.2

Step 3: Square each difference

-2.8*-2.8 = 7.84

-1.8*-1.8 = 3.24

-0.8*-0.8 = 0.64

1.2*1.2 = 1.44

4.2*4.2 = 17.64

Step 4: Calculate the average of the squared numbers to get the variance

7.84+3.24+0.64+1.44+17.64 = 30.8

30.8/5 = 6.16 (Variance)

Step 5: Find the square root of the variance

The square root of 6.16 = 2.48

Thus, Standard deviation of values 1,2,3,5 and 8 is 2.48

Activity 1.4

Read how standard deviation is used in calculating average rainfall in your city.

Graphically, the standard deviation 2.48 can be represented like below:

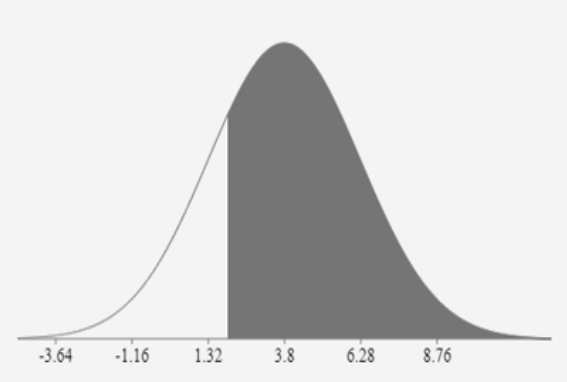


Fig 1.16 *Standard Deviation Curve*

Few real-life implementations of standard deviation include:

1. Grading Tests – If teacher wants to know whether students are performing at the same level or

whether there is a higher standard deviation.

2. To calculate the results of any Survey – If someone wants to have some measure of reliability of the responses received in the survey, to predict how bigger group of people may answer the same questions.
3. Weather Forecasting – If a weather forecaster is analyzing the low temperature forecasted for three different cities. A low standard deviation will always show reliable weather forecast.

10. Activity

What is the average family size of households of each student in your school?

Consider that your school head wants to calculate the average family size of students in your school. To carry out this activity at a large scale, let us first break it down into smaller sub parts.

Thus, to start with, we will ask teachers to take our family size of students of each classroom. We need to get an answer to the following questions over here:

1. What is the intended population? (Households of students in each teachers' class)
2. The variable to be measured. (The number of people in a household)
3. Anticipating variability. (Asking about typical household sizes)

Let us now move ahead and start getting answers to these questions step by step.

Collect/consider data

Suppose the teacher decides to work with five students at a time in the classroom and asks each student, "How many people, including yourself, are in the household that you live in?". As an answer to this, each student represents their family size with a collection of snap cubes.

The data for "family size" is represented with snap cubes in *Fig 1.17*

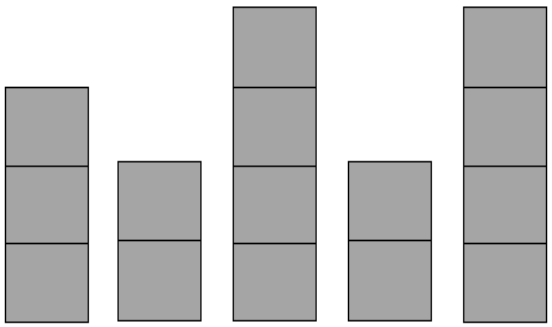


Fig 1.17 *Snap cube stacks representing family size*

Analyze the Data

To examine the distribution of the household size of the collected data, students first need to arrange the stacks of snap cubes in increasing order as shown in *Fig 1.18.*

You must have realized that family sizes vary. The next question that we need to ask is, *How many people would be in each family if all five families were the same size?*

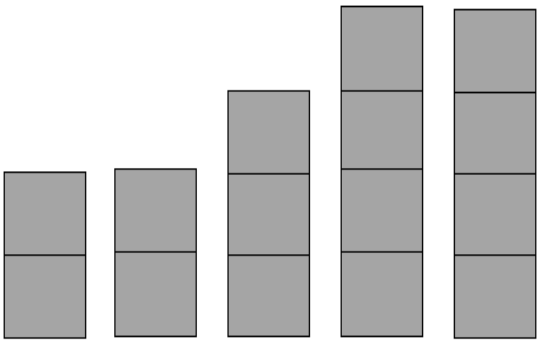


Fig 1.18 *Ordered stacks representing family size*

When we make all the family sizes the same, family size does not vary. You may use two equivalent approaches:

1. Disconnect all the snap cubes and redistribute them one at a time to the five students until all snap cubes have been allocated. In this case, there are 15 snap cubes. Redistributing them among the five students yields 5 stacks of 3 cubes each.
2. Remove one snap cube from the highest stack and place it on one of the lowest stacks, continuing until all the stacks are leveled out.

Both these approaches yield an equal family size of three, which we can consider as an equal share or a fair share.

For the second approach, you can start with removing a snap cube from the highest stack and placing it on one of the lowest stacks. This will result in a new arrangement of cubes as shown in *Fig 1.19*

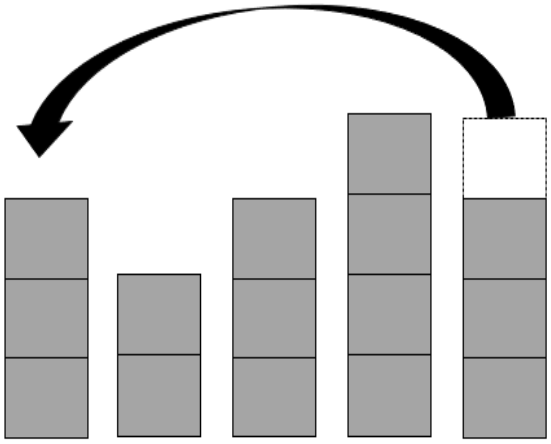


Fig 1.19 *Moving one snap cube from the highest stack*

We will continue this process until all the stacks are level, or nearly level when there is a remainder as shown in *Fig 1.20.*

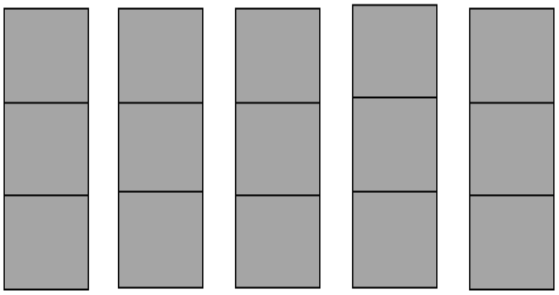


Fig 1.20 *Snap cubes representing an average family size*

After the final move all five stacks are levelled with three cubes each. This represents that the family size of three is an equal share. That means, if all five family sizes were the same, the number of people in the household would have been three. This equal share is nothing but the **mean** of the distribution.

By now, we know how to calculate the mean by adding up all the observations

and dividing it by the number of observations. However, what does mean tell us about the distribution? How are we expected to interpret the mean? How are we expected to describe the variability in a distribution in relation to its mean?

We can investigate the following problem to get an answer to these questions:

Suppose two other groups of five students in the classroom found their equal share value to be six. What are some different snap cube representations that they could have constructed?

To answer this, we should first realize that we need to start with 30 snap cubes. We can then create two different distributions of family size where the equal share value is 6. For example, consider the following two groups, Group 1 (shown in *Fig 1.21*) and Group 2 (shown in *Fig 1.22*) of data on five family sizes from the classroom where the equal share family size for each group is 6.

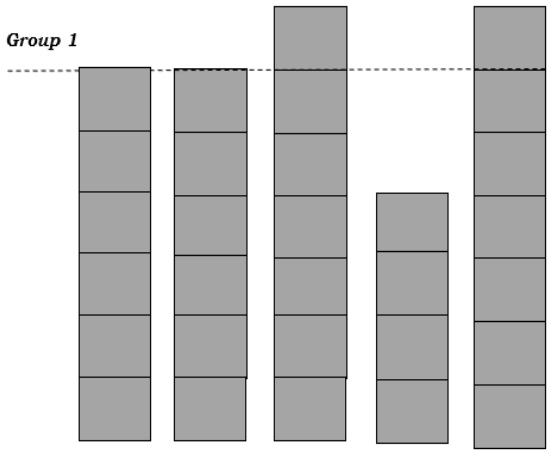


Fig 1.21 *Group 1 arrangement with average 6*

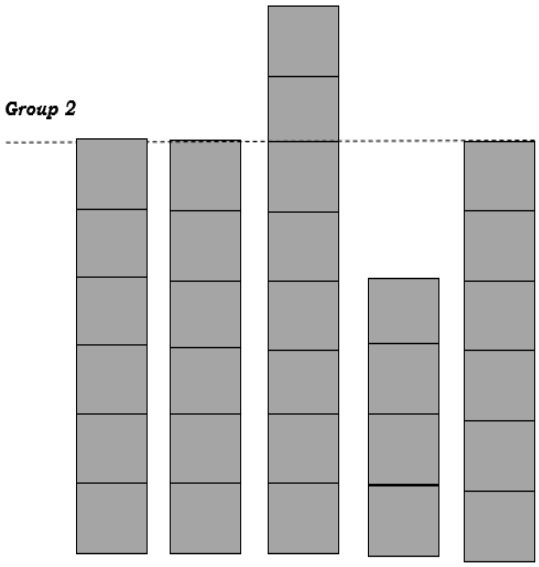


Fig 1.22 *Group 2 arrangement with average 6*

Because the equal share value for each group is 6, the two groups cannot be distinguished based on the equal share value. An analysis question in this case may be:

Which group is closer to being equal?

We can offer different answers to this question, including:

1. Group 2, as this group has the highest frequency of stacks of six snap cubes.
2. Group 1, as for this group, we need fewer snap cubes to level out all the stacks to the equal share value of six.

The second method of having fewer snap cubes to move can be thought of as counting the "number of steps to equal", or, how many steps we need to move the snap cubes to create the equal-sized

groups. Fewer steps indicate that the distribution is closer to being equal and has less variability from the mean. We can go through the process to check that for Group 1, we need to move two cubes a total of two steps. For Group 2, we need to move two cubes a total of two steps each. Thus, Group 1 and Group 2 has equal variability from the mean.

Interpret the results

Now, it is time to interpret the results to answer the original question,

What is the average family size of students in your school?

Using the results from the last two groups, we can comment that if the families are of equal size, the number of people in a household will be six. This will be **equal share or mean value**.

Thus, with the help of this activity we have learnt how mean helps us to get a quick resolution to our day to day activities.

Recap

- Sub setting is used to get a smaller chunk of data from a big data set
- A two-way table is a statistical table that shows the observed number or frequency for two variables
- Mean is a measure of central tendency that indicates the average value of a data set
- Median is the middle point of a sorted data set.
- Mean Absolute Deviation is the average of how far away all values in a data set are from the mean

Exercises

Objective Type Questions

Please choose the correct option in the questions below.

1. We want to get the cars of red color from the below data set. Which type of subsetting should be used?

Name	Height	Color
Innova	70	White
Swift	50	Red
Amaze	50	Red
Bolero	80	Gray

   a) Column based subsetting
   b) Data based subsetting
   c) Row based subsetting
   d) None of the above

2. Which is a more accurate measure of central tendency when there are outliers in the data set?
   a) Mean
   b) Median

3. Mean absolute deviation is an identifier of the variability of the data set. Is this a correct statement?
   a) Yes
   b) No

4. The mean absolute deviation is divided by coefficient of mean absolute deviation to calculate
   a) Variance
   b) Median
   c) Arithmetic Mean
   d) Coefficient of Variation

5. In a manufacturing company, the number of employees in unit A is 40, the mean is Rs. 6400 and the number of employees in unit B is 30 with the mean of Rs. 5500 then the combined arithmetic mean is

a) 9500
b) 8000
c) 7014.29
d) 6014.29

6. The mean deviation about the mean for the following data:
   5, 6, 7, 8, 6, 9, 13, 12, 15 is:
   a) 1.5
   b) 3.2
   c) 2.89
   d) 5

7. The arithmetic mean of the numerical values of the deviations of items from some average value is called the
   a) Standard Deviation
   b) Range
   c) Quartile Deviation
   d) Mean Deviation

Standard Questions

1. Explain the different ways of subsetting data.
2. When should we use median over mean?
3. What is Mean Absolute Deviation?
4. What is a two way relative frequency table? How is it different from two way frequency table?
5. What are two way frequency table beneficial for?
6. What is Standard Deviation?
7. How to calculate Standard Deviation?
8. Name five real-life applications of Standard Deviation
9. Explain five real-life situations where subsetting data can be advantageous

Higher Order Thinking Skills (HOTS)

1. Draw a graph to represent Standard deviation of 4.6
2. Calculate the mean of this data set - [56, 89, 76, 58, 58, 65]
3. Calculate the median of this data set - [56, 89, 76, 58, 58, 65]

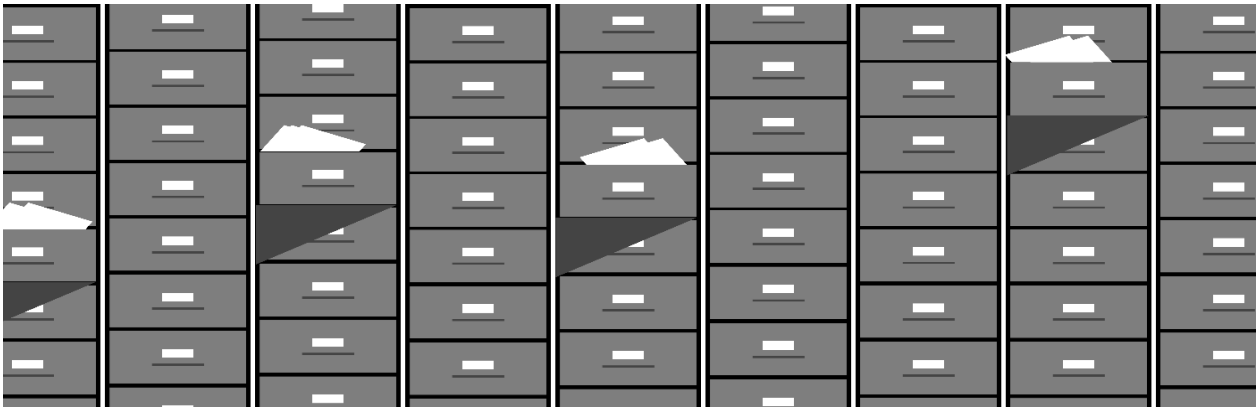
Applied Project

Calculate the average student height and weight for students in your classroom.

CHAPTER

2

DISTRIBUTIONS IN DATA SCIENCE



Studying this chapter should enable you to understand:

- What is distribution in data science?
- Different types of continuous distribution
- Different types of discrete distribution

1. Introduction

Now that we have understood the statistical terminology that are frequently used in data science in previous chapter, it is now time to learn about distribution of data in statistics. In this chapter, we will learn about different types of data distributions and characteristics of each distribution in detail.

2. What is distribution in data science?

Distribution in data science is a method which shows the probable values for a variable and how often they occur.

While the concept of probability gives us the mathematical calculations, distributions help us actually visualize what is happening underneath.

For example, consider a coin which has two sides, head and tail. Now when you throw the coin up in the air, what is the probability of getting a head? It is ½ or half right? And what is the probability of getting a tale? It is again ½ or half. However, if you say, what is the probability of getting a 3rd side on coin? Isn't it NIL? It is impossible to get a third side on the coin which has only two sides, head and tale. Thus probability is zero.

The distribution of an event consists not just the input values that can be seen, but made up of all possible values.

So, the distribution of the event, tossing the coin will be given by the following table. The probability of getting the head is 0.5. The probability of getting the tail is 0.5 and so on. You can be sure that you have exhausted all the values when the sum of probabilities is equal to 1% to 100%. For all other values apart from this, the probability of occurrence is zero.

Outcome	Probability
Head	0.5
Tail	0.5
All else	0

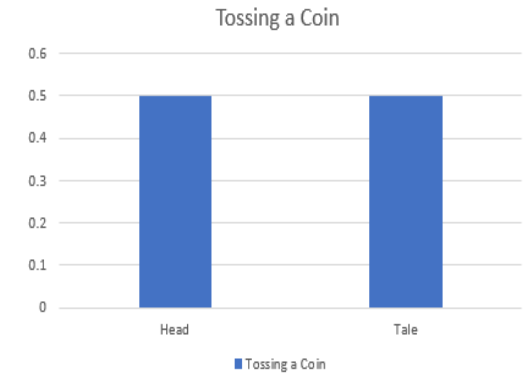


Fig 2.2 Uniform Distribution Graph for Tossing a Coin

Every probability distribution is associated with a graph which describes the likelihood of occurrence of each event. Below graph represents our example. This type of distribution is called as a **Uniform Distribution**.

However, point to note over here is that distribution in statistics is defined by underlying probabilities and not by the graph. Graph is just a visual representation. We had studied topic of data visualization in the previous grade.

Now, let us extend our problem statement to tossing two coins. What are the possibilities over here? Head-Head, Head-Tail, Tail-Tail and Tail-Head. Below is the table of all possible combinations.

Outcome	Probability
Head-Head	0.25
Head-Tail	0.25
Tail-Head	0.25
Tail-Tail	0.25
All else	0

***Fig 2.3** Probability Table for Tossing two coin*

Let us now understand the probability distribution for this scenario. Look at the below graph.

By looking at the graph we can understand that probability of getting a head in both the coins is 0.25. Similarly, getting a head in one coin and tail in another coin is 0.25. Probability of getting tail in one coin and head in another coin is 0.25. And probability of getting a tail in both the coins is 0.25.

Fig 2.4 Uniform Distribution Graph for Tossing two Coins

Thus, the graph of probability distribution in this case should look as shown in *Fig 2.4*

3. What are different types of distributions?

Types of distributions in data science is solely based on what kind of data we can encounter with while dealing with problems.

The data can be discrete or continuous.

Discrete Data is the data that takes only specified values. For example, if you give a test, you can either pass or fail. So, data is discrete in this case as it has only two specified outcomes.

Continuous Data is the data that can take any value within a given range. This range can be either finite or infinite. For example, depth of an ocean, weight of a person or length of a road.

Activity 1.2

Read about income distribution between different classes of people in India and try to classify it in a type of distribution

4. Statistical Problem Solving Process

The purpose of Statistical Problem-Solving Process which is detailed in *Fig 2.5* is to collect and analyze data to answer the statistical investigative questions.

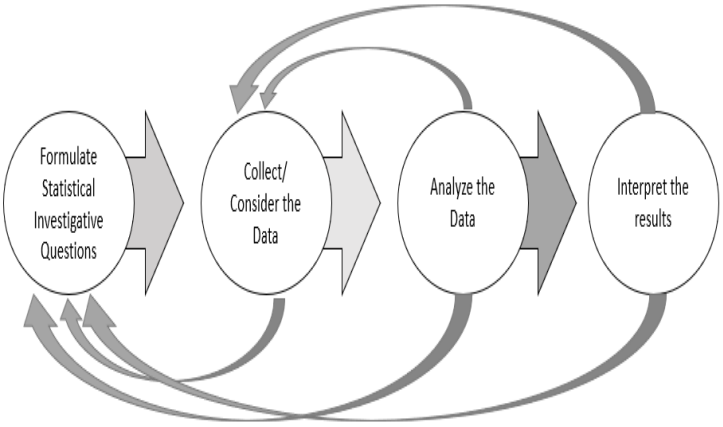


***Fig 2.5** Statistical Problem Solving Process*

This investigative process involves four components, each of which involves exploring and addressing variability:

1. Formulate Statistical Investigative Questions
2. Collect/Consider the Data
3. Analyze the Data
4. Interpret the Data

Let us understand each step-in detail now.

Formulate Statistical Investigative Questions

This can also be called as anticipating variability while beginning with the process.

Formulating statistical investigative questions that anticipate variability leads to productive investigations. For example, the following are all statistical investigative questions that anticipate variability and can lead to a rich data collection process and subsequent analysis of the data:

- How fast can my plant grow?
- Do plants exposed to more sunlight grow faster?
- How does sunlight affect the growth of a plant?

In contrast, the question *How tall is the plant?* is answered with a single height, it is therefore not a statistical investigative question.

How tall is the plant is a question we ask to collect data? Many other data collection questions could be asked to help collect the necessary data to answer the statistical investigative question: *Do plants exposed to more sunlight grow faster?* The fact that there will be differing heights for the different exposures of sunlight implies that we anticipate an answer based on measurements of plant heights that vary. While statistical investigative questions begin worthwhile studies, the use of questioning is prominent throughout all four components of the statistical problem-solving process. Such uses of questioning will be illustrated throughout the examples at the different levels.

In addition to anticipating variability, there are other features of a statistical investigative question that are important. The variables of interest must be clear, the group or population that the question is focused on must be clear, the intent of the question should be clear – is the question requiring a description of the data, is the question comparing a variable across two or more groups, is the question looking at an association between two variables, the question should be about the whole group (anticipating variability) and not about an individual (giving a deterministic answer), the question should be answered through data collection (primary data) or with the data in hand (secondary data), and the question should be purposeful.

Collect/Consider the Data

This step can be called as acknowledging variability while designing for differences.

Data collection designs must acknowledge variability in data. Few methods are used to reduce and detect variability in data, such as Statistical Process Control and random sampling. While others are used to induce variability to test treatments, such as Design of Experiments. In the latter approach, experimental designs are chosen to acknowledge the differences between groups subjected to different treatments. Random assignment to the groups is intended to reduce differences

between the groups due to factors that are not manipulated or controlled in the experiment. In all designs, the main statistical focus is to look for, account for and explain variability.

After the data is available, whether it is collected first-hand or acquired from another source, it needs to be interrogated. For example, questions about how the variables differ by type, the possible outcomes of each of the variables, and how the data was collected is necessary to clarify whether the data is useful for answering the statistical investigative question. The data collection design impacts the scope of generalizability and the possible limitations in analysis and interpretation.

Analyze the Data

This step can also be called as accounting of variability while the distributions.

When we analyze the data, we try to understand its variability. Reasoning about distributions is key to accounting for and describing variability at all development levels. Graphical displays and numerical summaries are used to explore, describe, and compare variability in distributions.

For example, the batting averages of Indian Cricket Team and the batting averages of Australia Cricket Team for a particular year can be displayed in two comparative dot plots and boxplots. These graphs show the variability of each teams' distributions of batting averages. We can consider variability by describing the overlap and the separation of the distribution of the two teams.

Interpret the Results

This step can also be called as allowing for variability while looking beyond the data.

Statistical interpretations are made in the presence of variability and must take variability into account. When interpreting the results of a randomized comparative medical experiment, we must remember there are two important sources of variability: randomization to treatment group, and variability from individual to individual. When we generalize the results and look beyond the study data collected; we must consider these sources of variability.

5. Activity – Choosing groups for school dance program

Consider that there is an annual event in your school for which you all are planning to shortlist a musical group for a single grade. You can do this by conducting a class census. Let us ask following statistical question to go start with the activity:

What type of music do the students in our grade like?

To start with, we can start collecting data for each class. We will collect data for entire population of the class i.e. each student will answer this question. Later, we will extend the collection to the entire grade.

On similar grounds, we may also plan to collect data for the entire school. This is because one single class may not represent the preferences of all students in one grade or all students at the school.

Now that we have all the data with us, each class can compare preferences of their class with the preferences of other classes of the school and explore the following statistical question:

What type of music do the students at our school like?

Collect/consider data

The statistical question,

What type of music do the students at our school like?

asks about the preferences of students at the school overall. In this case, a data collection plan may use a single class, for example, grade 10 English class, as a sample to make decisions for the whole school. For this situation, we can discuss the limitations of the chosen sample. Alternatively, what we can do is, we can randomly select few students from each class or select couple of classes and get all the students in those classes to complete a survey.

To excel further, we can improve on survey questions used before by understanding potential pitfalls to avoid in survey design (like ambiguous wording and misleading questions) or maybe by providing more choices in the answers. Additionally, we can collect the data on multiple aspects of the topic which can foreshadow answering the other statistical investigative questions.

For example, we can pose a series of survey questions that allow us to explore in more depth the types of music students like. After collecting all the data, we can look at whether an association appears to be likely between different types of music students like. This information might tell us the choice of music for the school dance.

Question 1: Tick yes for any of the following music types you like. Check no for any you don't like

	Yes	No
Classical	☐	☐
Rap	☐	☐
Bollywood	☐	☐
Pop	☐	☐
Rock	☐	☐

Question 2: What is your most favourite type of music?

- ☐ Classical
- ☐ Rap
- ☐ Bollywood
- ☐ Pop
- ☐ Rock

Question 3: What is your second favourite type of music?

- ☐ Classical
- ☐ Rap
- ☐ Bollywood
- ☐ Pop
- ☐ Rock

Question 4: Would you prefer a live band or a DJ at the annual dance event at school?

- ☐ DJ
- ☐ Live Band

Analyze the data

Many of the graphical, tabular, and the numerical summaries that we have made while data collection can be enhanced and used for more sophisticated analyses at later stages. Displays at later stages mostly represent multiple variables and/or use multiple displays to answer the statistical investigative questions. To analyze the survey data collected using a class as a sample for the school, we can graph the number of students who like each type of music. The bar graph in *Fig 2.6* uses the student answers to survey question 1 noted above, where each music type is a variable.

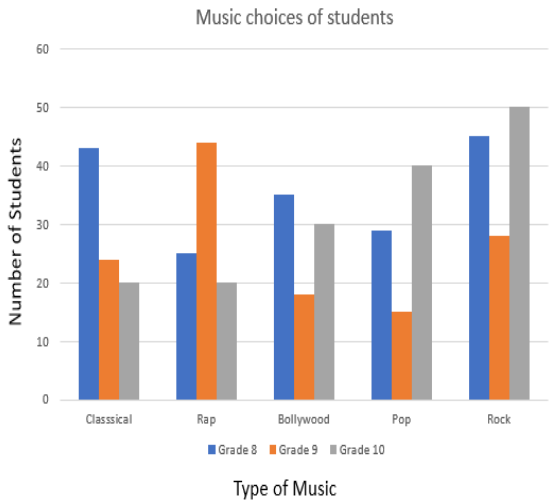


Fig 2.6 *Side by side bar graph for liking music*

The bar graph shows the frequencies of students who like and dislike each type of music. From this graph, it is evident that Rock has the highest number of students in the class responding to yes, that they like it. The second highest is Rap followed by Classical. Thus, the graph suggests that Rock, Rap and Classical are the most liked music among students.

Responses to survey question 2 can be analyzed to see what students favourite type of music is. *Fig 2.7* shows there are an equal number of students who voted Classical and Rock as their favorite music.

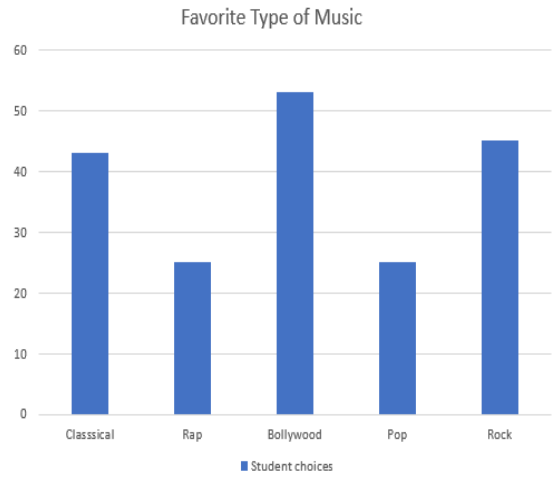


Fig 2.7 *Favorite type of music*

Students can explore favorite and second favorite music which are answers to questions 2 and 3 on the survey through two-way graph as shown in *Fig 2.8*.

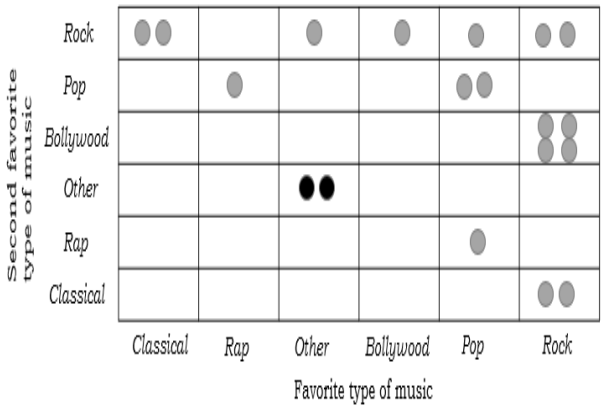


Fig 2.8 *Favorite and second favorite type of music*

Here, they can say that all but two students have Pop and Rock as their first and/or second choices for type of music preferences. In this graph, each dot represents a student in the class who responded to the survey. This two-way graph shows 36 bins (6 possible second favorite music types by 6 possible favorite music types). The bin in the top left corner has two dots in it, representing the two students in the class who answered that their favorite type of music is Classical, and their second favorite type of music is Rock.

Analysis of the favorite and second favorite types of music shows that nearly all the students (17 students, those shaded lighter in *Fig 2.8*) in the class have voted Bollywood, Pop and Rock as their first or second choice. Only two students (those shaded in a darker color) did not rank Bollywood, Pop and Rock in top two.

Students can also look at the choice between live band and DJ and add this to their final conclusions about the types of music that students like (*Fig 2.9*). The difference in the number of students who prefer a live band versus a DJ is zero.

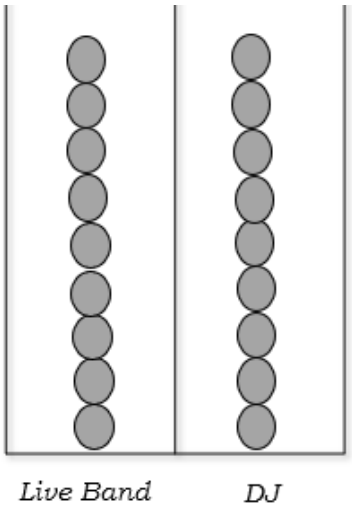


Fig 2.9 *Choice of live band or DJ*

Interpret Results

The analysis shows that Bollywood, Pop and Rock are very similar in terms of the number of students who choose them as their favorite type of music. Most of the students picked up Bollywood, Pop, and Rock as their favorite and/or second favorite type of music. In addition, students might note that a live band and a DJ are equally chosen among class members.

Recap

- Distribution in data science is a method which shows the probable values for a variable and how often they occur.
- Discrete Data is the data that takes only specified values.
- Continuous Data is the data that can take any value within a given range. This range can be either finite or infinite.

Exercises

Objective Type Questions

Please choose the correct option in the questions below.

1. If a card is chosen from a standard deck of cards, what is the probability of getting a five or a seven?
   a) 4/52
   b) 1/26
   c) 8/52
   d) 1/169

2. Which of the following is the condition for Uniform Distribution?
   a) Each value in the set of possible values has the exact same possibility of happening.
   b) Have a constant probability of success
   c) Has only two possible outcomes
   d) Must have at least 3 trials

3. The collection of one or more outcomes from an experiment is called

a) Probability
b) Distribution
c) Event
d) Random Experiment

4. Which of the following are types of distributions?
   a) Continuous
   b) Discrete
   c) Both of them

5. Which of the following is not an example of discrete probability distribution?
   a) The sale or purchase price of a house
   b) The number of bedrooms in a house
   c) The number of bathrooms in a house
   d) Whether or not a home has a swimming pool in it

6. A discrete probability distribution may be represented by
   a) A table
   b) A graph
   c) A Mathematical Equation
   d) All of these

7. What is the probability that a ball is drawn at random from a jar?
   a) 0.1
   b) 1
   c) 0.5
   d) 0
   e) Cannot be determined from given information

8. Statistical investigative process has which of the following components:
   a) Formulate Statistical Investigative Questions
   b) Collect/Consider the Data
   c) Analyze the Data
   d) Interpret the Data
   e) All of the above

Standard Questions

1. Explain what distribution in data science with the help of two examples is.
2. Explain what is a Statistical Problem-Solving process.
3. Explain how distributions are broadly categorized. Support your answer with appropriate examples for each category.
4. Explain in detail how do we formulate statistical investigative questions.
5. Name five instances where you have observed a uniform distribution.

Higher Order Thinking Skills (HOTS)

1. Consider that there are 60 students in your class out if which 20 get affected with cold and flu every semester. Note down five statistical investigative questions for determining a students' immunity to a catching cold and flu.

2. Consider you are taking a part in an animal welfare campaign. One of the most recent concerns raised by people is dogs not being able to tolerate sudden rise in temperatures due to global warming. Note down five statistical investigative questions to understand how dogs react to changing weather.

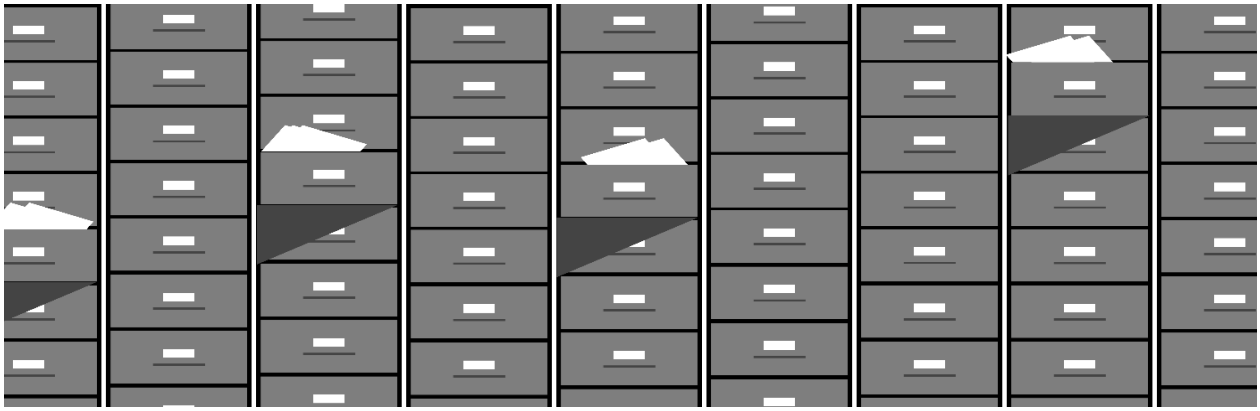
Applied Project

Consider that you have a food event in your residential society. Perform detailed analysis and interpret what should be the top five cuisines that most people in the society prefer for this event.

CHAPTER

3

IDENTIFYING PATTERNS



Studying this chapter should enable you to understand:

- How to identify partiality, preference and prejudice
- What is Central Limit Theorem?

1. What is partiality, preference and prejudice?

We often come across situations where if we have a special fondness towards a particular thing, we tend to be slightly partial towards it. This, in majority cases may affect the outcome or you can say it can deviate the outcome in favor of certain thing. Naturally, it is not the right way of dealing with the data on larger scale.

This partiality, preference and prejudice towards a set of data is called as a **Bias**.

In Data Science, bias is a deviation from the expected outcome in the data. Fundamentally, you can also call bias as error in the data. However, it is observed that this error is indistinct and goes unnoticed. So, question to be asked is, why does the bias occur in first place?

Bias basically occurs because of sampling and estimation. If we would know everything about all the entities in our data and would store information on all probable entities, our data would never have any bias. However, data science is often not conducted in carefully controlled conditions. It is mostly done of the "found data", i.e. the data that is collected for a purpose other than modelling. That is the reason why this data is very likely to have biases.

Next question that may arise in your mind is, why does the bias really matter? Well, the answer is that predictive

models often consider only the data that is used for training. In fact, they know no other reality other than the data that is fed in their system. Naturally, if the data that is fed into the system is biased, model accuracy and fidelity are compromised. Biased models can also tend to discriminate against certain groups of people. Therefore, it is very important to eliminate the bias to avoid these risks.

2. How to identify the partiality, preference and prejudice?

We can categorize common statistical and cognitive bias in following ways:

1. Selection Bias
2. Linearity Bias
3. Confirmation Bias
4. Recall Bias
5. Survivor Bias

Selection Bias

This type of bias usually occurs when a model itself influences the creation of data that is used to train it. Selection bias is said to occur when the sample data that is gathered is not representative of the true future population of cases that the model will see. This bias occurs mostly in systems that rank the content like recommendation systems, polls or personalized advertisements. This is because, user responses for the content that is displayed is collected and the user response for the content that is not displayed is unknown.

Linearity Bias

Linearity bias assumes that change in one quantity produces an equal and proportional change in another. Unlike selection bias, linearity bias is a cognitive bias. This is produced not through some statistical process but rather through how mistakenly we perceive the world around us.

Confirmation Bias

Confirmation Bias or Observer Bias is an outcome of seeing what you want to see in the data. This can occur when researchers go into a project with some subjective thoughts about their study, which is either conscious or unconscious. We can also encounter this when labelers allow their subjective thoughts to control their labeling habits, which results in inaccurate data.

Recall Bias

Recall Bias is a type of measurement bias. It is common at the data labeling stage of any project. This type of bias occurs when you label similar type of data inconsistently. Thus, resulting in lower accuracy. For example, let us say we have a team labeling images of damaged laptops. The damaged laptops are tagged across labels as damaged, partially damaged, and undamaged. Now, if someone in the team labels an image as damaged and some similar image as partially damaged, your data will obviously be inconsistent.

Survivor Bias

The survivorship bias is based on the concept that we usually tend to twist the data sets by focusing on successful examples and ignoring the failures. This type of bias also occurs when we are looking at the competitors. For example, while starting a business we usually take the examples of businesses in a similar sector that have performed well and often ignore the businesses which have incurred heavy losses, gone bankrupt, merged etc.

While this is arguable point that we don't want to copy the failure, we can still learn a lot by understanding a range of customer experiences. The only way to avoid survivor bias in our systems is by finding as many inputs as possible and study the failures as well as average performers.

3. Probability for Statistics

Probability is all about counting randomness. It is the basics of how we make predictions in statistics. We can use probability to predict how likely or unlikely particular events may be. We can also, if needed, consider informal predictions beyond the scope of the data which we have analyzed.

Probability is a very essential tool in statistics. There are two problems and nature of their solution that will illustrate the difference.

Problem 1: Assume a coin is "fair"

Question: If a coin is tossed 10 times, how many times will we get "tail" on the top face.

Problem 2: You pick up a coin

Question: Is this a fair coin? That is, does each face have an equal chance of appearing?

Problem 1 is a mathematical probability problem. Problem 2 is a statistics problem that can use the mathematical probability model determined in Problem 1 as a tool to seek a solution.

The answer to neither question is deterministic. Tossing coin produces random outcomes, which suggests that the answer is probabilistic. The solution to Problem 1 starts with the assumption that the coin is fair. It later proceeds to logically deduce the numerical probabilities for each possible count of "tails" after a toss resulting from 10 tosses. The possible counts are $0, 1 \ldots, 10$.

The solution to Problem 2 starts with an unfamiliar toss; we do not know if it is fair or biased. The search for an answer is experimental: toss the coin, see what happens, and examine the resulting data to see whether they look as if they came from a fair toss or a biased toss. One possible approach to making this judgement would be the following: Toss the coin 10 times and record the number of count when you got a "tail". Repeat this process of tossing the coin 100 times. Compile the number of times you got a "tail" in each of these 100 trials. Compare these results to the frequencies produced by the

mathematical model for a fair toss in Problem 1. If the frequencies from the experiment are quite dissimilar from those predicted by the mathematical model for a fair toss and the observed frequencies are not likely to be due to chance variability, then we can conclude that the toss is not fair.

In Problem 1, we form our answer from logical deductions. In Problem 2, we form our answer by observing experimental results.

4. The Central Limit Theorem

The Central Limit Theorem states that distribution of sample approaches a normal distribution as the sample size gets larger irrespective of what is the shape of the population distribution.

The Central Limit Theorem is a statistical theory stating that given a significantly large sample size from a population with finite variance, the mean of all samples from same set of population will be roughly equal to the mean of the population. This holds true regardless of whether the source population is normal or skewed provided that the sample size is significantly large.

Few points to note about the Central Limit Theorem are:

- ✓ The Central Limit Theorem states that the distribution of sample means nears a normal distribution as the sample size gets bigger.
- ✓ Sample sizes that are equal to or greater than 30 are considered enough for the Central Limit Theorem to hold.
- ✓ Key aspect of the Central Limit Theorem is that the average of sample mean, and the standard deviation will always equal the population mean and the standard deviation.
- ✓ A significantly large sample size can predict the characteristics of a population very accurately.

Let us now understand the Central Limit Theorem with the help of an example.

Consider that there are 50 houses in your area. And each house has 5 people. Our task is to calculate average weight of people in your area.

The usual approach that majority follow is:

1. Measure the weights of all people in your area
2. Add all the weights
3. Divide the total sum of weights with the total number of students to calculate the average

However, the question over here is, what if the size of data is enormous? Does this way of calculating the average make sense? Of course, the answer is no. Measuring weight of all the people will be a very tiring and lengthy process.

As a workaround, we have an alternative approach that we can take.

1. To start with, draw groups of people at random from your area. We will call this a sample. We will draw multiple samples in this case, each consisting of 30 people
2. Calculate the individual mean of each sample set
3. Calculate the mean of these sample means
4. To add up to this, a histogram of sample mean weights of people will resemble a normal distribution.

This is what the Central Limit Theorem is all about. Now let us move ahead and understand what the formula for the central limit theorem is.

$$\mu_{\bar{x}} = \mu$$

And,

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

Where,

μ = Population mean
σ = Population standard deviation
$\mu_{\bar{x}}$ = Sample mean
$\sigma_{\bar{x}}$ = Sample standard deviation
n = Sample size

Now, that we have understood what the central limit theorem is, let us now see what its real-life applications are and what is the formula to calculate it, let us learn why the central limit theorem is so important.

Let us have a look at below example of the Central Limit Theorem Formula:

Example:

In India, the recorded weights of the male population are following a normal distribution. The mean and the standard deviations are 68 kgs and 10 kgs, respectively. If a person is eager to find the record of 50 males in the population, then what would mean and the standard deviation of the chosen sample?

Over here,

Mean of Population – 68 kgs

Population Standard Deviation (σ) – 10 kgs

Sample size (n) – 50

Solution:

Mean of Sample is the same as the mean of population.

The mean of the population is 68 since the sample size > 30.

Sample Standard Deviation is calculated using below formula:

$\sigma_x = \sigma/\sqrt{n}$

Thus, Sample Standard Deviation = $10/\sqrt{50}$

Sample Standard Deviation is **1.41.**

5. Why is the Central Limit Theorem important?

The Central Limit Theorem states that no matter what the distribution of population is, the shape of the sampling distribution will always approach normality as the sample size increases.

This is helpful, as any research never knows which mean in the sampling distribution is the same as population mean, however, by selecting many random samples from population, the sample means will cluster together, allowing the researcher to make a good estimate of the population mean.

Having said that, as the sample size increases, the error will always decrease.

Some practical implementations of the Central Limit Theorem include:

1. Voting polls estimate the count of people who support a particular election candidate. The results of news channels that come with confidence intervals are all calculated using the Central Limit Theorem.
2. The Central Limit Theorem can also be used to calculate the mean family income for a specific region.

Activity 3.1

Read about how population is measured for your city and how the Central Limit Theorem can help in counting the large group of population.

Recap

- In Data Science, bias is a deviation from the expected outcome in the data.
- Selection bias is said to occur when the sample data that is gathered is not the representative of the true future population of cases that the model will see.
- Linearity bias assumes that change in one quantity produces an equal and proportional change in another.
- Confirmation Bias or Observer Bias is an outcome of seeing what you want to see in the data.
- Recall bias occurs when you label similar type of data inconsistently.
- The survivorship bias is based on the concept that we usually tend to twist the data sets by focusing on successful examples and ignoring the failures.
- The Central Limit Theorem states that distribution of sample approaches a normal distribution as the sample size gets larger irrespective of what is the shape of the population distribution.

Exercises

Objective Type Questions

Please choose the correct option in the questions below.

1. What is the Data Science term used to describe partiality, preference, and prejudice?
   a) Bias
   b) Favoritism
   c) Influence
   d) Unfairness

2. Which of the following is NOT a type of bias?
   a) Selection Bias
   b) Linearity Bias
   c) Recall Bias
   d) Trial Bias

3. Which of the following is not a correct statement about a probability
   a) It must have a value between 0 and 1
   b) It can be reported as a decimal or a fraction
   c) A value near 0 means that the event is not likely to occur/happen
   d) It is the collection of several experiments

4. The central limit theorem states that sampling distribution of the sample mean is approximately normal if
   a) All possible samples are selected
   b) The sample size is large
   c) The standard error of the sampling distribution is small

5. The central limit theorem says that the mean of the sampling distribution of the sample mean is
   a) Equal to the population mean divided by the square root of the sample size
   b) Close to the population mean if the sample size is large
   c) Exactly equal to the population mean

6. Sample of size 25 are selected from a population with mean 40 and standard deviation 7.5. The mean of the sampling distribution sample mean is
   a) 7.5
   b) 8
   c) 40

Standard Questions

1. Explain what is Bias and why it occurs in data science?
2. Explain Selection Bias with the help of an example
3. Explain Recall Bias with the help of an example
4. Explain Linearity Bias with the help of an example
5. Explain Confirmation Bias with the help of an example
6. What is the central limit theorem?
7. What is the formula for central limit theorem?
8. What is real life application of central limit theorem?
9. Why central limit theorem is important?
10. The coaches of various sports around the world use probability to better their game and create gaming strategies. Can you explain how probability is applied in this case and how does it help players?

Higher Order Thinking Skills

1. As per reports, in October 2019, researchers found that an algorithm used on more than 200 million people in US hospitals to predict which patients who would likely need extra medical care heavily favored white patients over black patients. Can you reason about what must have caused this bias and categorize it into the types of bias that you learnt in this chapter?

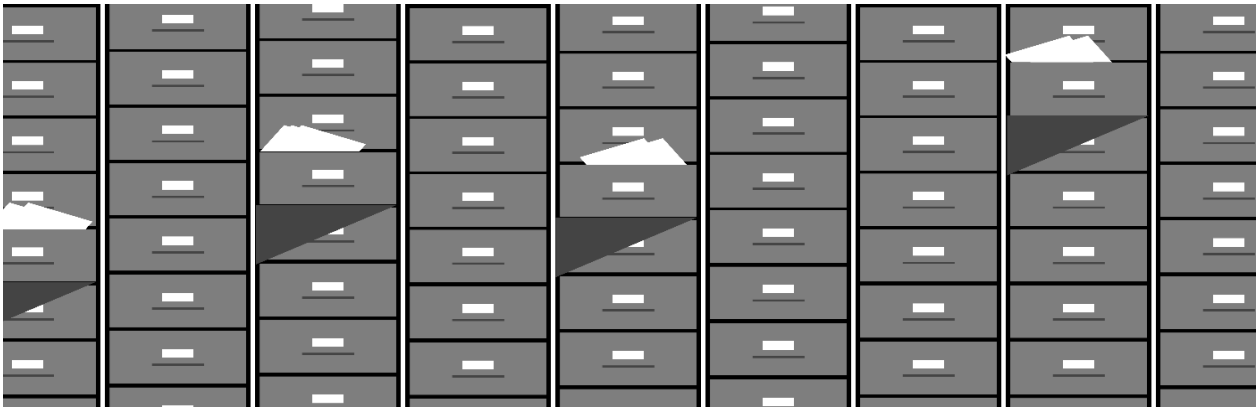
2. The recorded percentage of the population who speaks English in India are following a normal distribution. The mean and the standard deviations are 62 and 5, respectively. If a person is eager to find the record of 50 people in the population, then what would mean and the standard deviation of the chosen sample?

Applied Project

Consider that your friend is planning to open a clothing store in your area. With the help of central limit theorem, determine what should be the type and collection of clothes that will sell better in your area.

CHAPTER 4

DATA MERGING



Studying this chapter should enable you to understand:

1. How to merge data sets?
2. What is Standard Deviation and what are different ways to calculate it?

1. Overview of Data Merging

In Data Science, data merging is the process of combining two or more data sets into a single data frame. This process is necessary when we have raw data stored in multiple files or data tables, that we want to analyze all in one go.

However, while merging the data from different sources there are many issues that occur that require corrections for successful data merging. Different data sources will always have different naming conventions than the main data source. They may have different ways of grouping the data and so on. Many times, it happens that the additional data source happens to be created at a very different time by different people with a different objective and use-cases. Owing to all these factors, it should not sound strange if there is a lot of difference between multiple data sources.

In this topic, we will explore various ways of simplifying the process of data merging. There are many places where these data merging techniques will help you. For example, if you have two different systems that operate in parallel with each other. Suppose that you have to perform some analysis of the

relationship where you are having a legacy system with a very poorly formatted data that you are willing to integrate with your new system. This is where data merging comes into the picture. Let us now dive deep into data merging techniques.

We can perform data merging by implementing data joins on the databases in frame. There are three categories of data joins:

1. One to One Joins
2. One to Many Joins
3. Many to Many Joins

One to One Joins

One to one join is probably one of the simplest join techniques. In this type of join, each row in one table is linked to a single row in another table using a "key" column.

For example, in a company database, each employee has only one Employee ID, and each Employee ID is assigned to only one employee.

In the database, a one to one relationship looks like this:

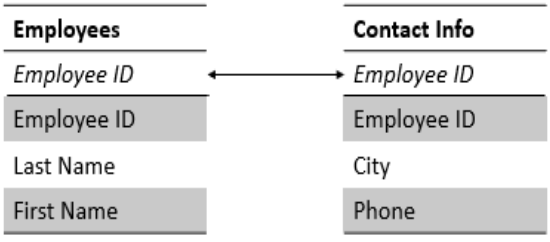


Fig 4.1 *Employee Table*

In this example, the "key" field in each table is "Employee ID". This "key" field is designed to contain unique values. In the Employee table, the Employee ID field is the primary key, in the Contact Info table, the Employee ID field is a foreign key.

The one to one relationship returns the related records when the value in the Employee ID field in the Contact Info table is the same as the Employee ID field in the Employees table.

This is how one to one join works, by merging the data tables using this primary key.

One to Many Joins

In a one to many join, one record in a table can be related to one or many records in another table. For example, each student can have multiple books by school library.

In the database, a one to many relationships looks like this:

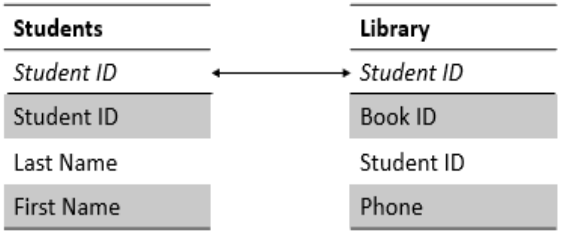


Fig 4.2 *Library Database*

In this example, the primary key field in the Students table, Student ID, is designed to contain the unique values. The foreign key field in the Library table, Student ID, is designed to allow multiple instances of the same value.

The one to many relationships returns the related records when the value in the Student ID field in the Library Table is the same as the value in the Student ID field in the Students table.

This is one to many join works, by merging databases using primary key which demonstrates one to many relationships.

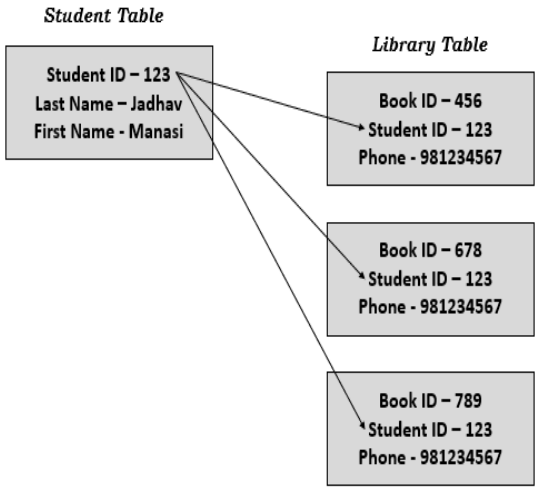


Fig 4.3 One to Many Relationship

Many to Many Joins

A many to many relationships is said to occur when multiple records in one table are related to multiple records of other table. For example, a many to many relationships exists between students and courses. A student can register for multiple courses. A course can have multiple students.

It is not easy to perform join on tables which have many to many relationships. As a workaround, to perform a join, you can break a many to many relationships into two one to many relationships by using a third table which is called as a join table. Every record in a join table contains a match field that contains the value of the primary keys of two tables that it joins. In a join table, usually these match fields are called as foreign keys. These foreign keys are populated with the data as records in the join table are created from either table that it joins.

The below table demonstrates the Student table, which contains a record for every student. It also contains a Courses table, which contains a record for each course. A join table called Enrollments creates two one to many relationships, the one between each of the two tables.

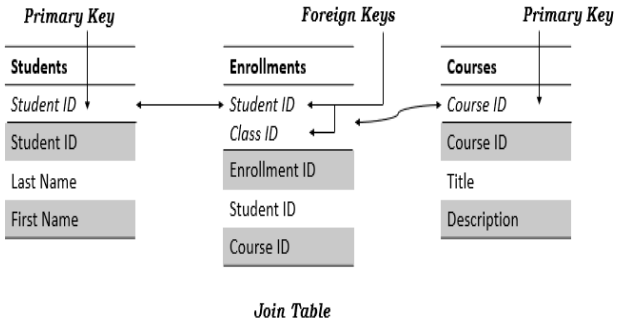


Fig 4.4 Students and Courses Database

The primary key Student ID is a unique identifier for every student in Students table. The primary key Course ID is a unique identifier for every course in the Courses table. The Enrollments table carries the foreign keys Student ID and Course ID.

To set up a join table for many to many relationships,

1. Using the above example, you can create a table called Enrollments. This will act as a join table.
2. In the Enrollments table, make a Student ID field and a Course ID field.
3. Make a relationship between the two Student ID fields in the tables. Later, make a relationship between two Course ID field in the tables.

We can use this design, if a student registers for four courses, we can ensure that the student has only one record in the Students table and four records in the Enrollments table, one for each course student is enrolled in.

2. What is Z-Score?

A Z-score describes the position of a point in terms of its distance from the mean when it is measured in the standard deviation units. The z-score is always positive if the value of z-score lies above the mean and it is negative if its value is below the mean.

Z-score is also known as standard score as it allows comparison of scores on different types of variables by standardizing the distribution.

A standard normal distribution is a normally shaped distribution with a mean of value as 0 and a standard deviation of value as 1.

3. How to calculate a Z-score?

The mathematical formula for calculating the z-score is as following:

$$\mathbf{Z = (x-\mu)/\sigma}$$

Where,

X = raw score

μ = Population mean

σ = Population Standard Deviation

Thus, the z-score is the raw score minus the population mean, divided by the population standard deviation.

Whenever we come across situations where the population mean and the population standard deviation are unknown, the standard score can be calculated using the sample mean i.e. $\bar{x}$ and the sample standard deviation as estimates of population values.

Now we will consider an example that will illustrate the use of z-score formula. Consider that we know about a population of group of kids having weights that are normally distributed. Further to this, consider that we know that the mean of the distribution is 10 kgs and the standard deviation is 2 kgs. Now consider the below questions:

1. What is the z-score for 12 kgs?
2. What is the z-score for 5 kgs?
3. How many kgs corresponds to a z-score of 1.25?

For the first question, we simply plug x=12 in our z-score formula. The result is: (12-10)/2 = 1.

This means that 12 is one standard deviations above the mean.

The second question is also very similar. Simply put x=5 into the formula. Thus, the result for this is:

(5-10)/2= -2.5

The interpretation of this is that 5 is 2.5 standard deviations below the mean.

For the last question, we now know our z-score. For this problem we plug z = 1.25 into the formula and use basic algebra to solve for x:

1.25 = (x-10)/2

Multiply both the sides by 2:

2.5 = (x-10)

Add 10 to both the sides:

12.5 = x

Hence, we see that 12.5 kgs corresponds to a z-score of 1.25.

4. How to interpret the Z-score?

The value of a z-score always tells us how many standard deviations we are away from the mean. For example, if the z-score is equal to 0, it is on the mean.

A positive z-score tells us that the raw score is higher than the mean average. For example, if the z-score is equal to +2, it is 2 standard deviations above the mean.

A negative z-score tells us that the score is below the mean average. For example, if a z-score is equal to -3, it is 3 standard deviations below the mean.

5. Why is a Z-score so important?

It is very helpful to standardize the values of a normal distribution by converting them into z-score because:

1. It gives us an opportunity to calculate the probability of a value occurring within a normal distribution.
2. Z-score allows us to compare two values that are from the different samples.

6. Concept of Percentiles

The maximum value of the distribution can be considered in an alternative way. We can represent it as a value in a set of data having 100% of the observations at or below it. When we consider the maximum value this way, it is called the 100^{th} percentile.

A percentile can be defined as the percentage of the total ordered observations at or below it. Therefore, p^{th} percentile of a distribution is the value such that p percentage of the ordered observation falls at or below it.

Consider the following data set: [10, 12, 15, 17, 13, 22, 16, 23, 20, 24]

Here, if we want to find the percentile for element 22, we follow the steps below:

- Sort the dataset in ascending order. Once sorted, the dataset will look like [10, 12, 13, 15, 16, 17, 20, 22, 23, 24]
- The number of values at or below the element 22 is 8. The total number of elements in the dataset is 10.
- Thus, going by the definition, 80 percent of the values are at or below the element 22. Thus, percentile for the element 22 is 80 percentiles.

7. Quartiles

Quartiles of dataset partitions the data into four equal parts, with one-fourth of the data values in each part. The total of 100% is divided into four equal parts: 25%, 50%, 75% & 100%. Since the median is defined as the middlemost value in the observation, the median will have 50% of the observations at or below it. Thus, the second quartile(Q_2) or the 50^{th} percentile demarcates the median. The most frequently used percentiles other than the median are the 25^{th} percentile and the 75^{th} percentile. The 25^{th} percentile defines the first quartile, the 75^{th} percentile defines the third quartile, and the 100^{th} percentile represents the fourth quartile.

The first quartile is the median of all the values to the actual median's (Q_2) left. Similarly, the third quartile is the median of all the values to the actual median's (Q_2) right.

Using the values of the quartiles, we can also find out the interquartile range. An **interquartile range** can be defined as the measure of middle 50% of the values when ordered from lowest to highest. The interquartile range can be calculated by subtracting first quartile(Q_1) from the third quartile(Q_3).

IQR = $Q_3 - Q_1$

Let us consider the following 10 data points:

[10, 20, 30, 40, 50,60, 70, 80, 90, 100]

Here, as there are ten values (an even number of values), the median is halfway between the fifth & sixth data values, which gives us 55 as the median, or Q_2.

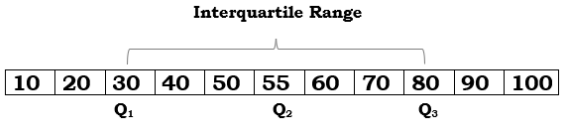
10	20	30	40	50	55	60	70	80	90	100
					Q_2					

The first quartile or Q_1 is the median of all the values to the left of Q_2. Thus here, 30 is the middle number of numbers to the left of the actual median (Q_2).

The third quartile or Q_3 is the median of all the values to the right of Q_2. Thus here, 80 is the middle number of numbers to the right of the actual median (Q_2).

10	20	30	40	50	55	60	70	80	90	100
		Q_1			Q_2			Q_3		

The interquartile range (IQR) can be calculated as $Q_3 - Q_1$, which is 80 - 30 = 50.



An important application of quartiles is in temperature ranges for the day as reported on a weather report. In the presence of irregularities, the range values can be significantly influenced by them. Hence, it is preferred to use the IQR instead, thereby ignoring the top 25 percentile and the bottom 25 percentile of the data points. In the presence of irregularities, IQR is more robust as well as a better representation of the amount of spread in the data.

8. Deciles

Just like quartiles, we have deciles. While quartiles sort the data into four quarters, deciles sort the data into ten equal parts: the 10^{th}, 20^{th}, 30^{th}, 40^{th}, 50^{th}, 60^{th}, 70^{th}, 80^{th}, 90^{th}, 100^{th}.

The higher the place in the decile ranking, the higher is the overall ranking. For example, a person receiving 99 percentiles in a test would be placed in a decile ranking of 10. However, a person receiving 5 percentiles in the same test would be placed in a decile ranking of 1.

The mathematical formula to calculate decile is:

$$Di = \frac{i * (n+1)}{10th\ Data}$$

Where ***n*** is the number of data in the population sample.

i is the i^{th} decile and can be represented as:

1^{st} Decile, $D_1 = 1 * (n + 1)/ 10^{th}$ data

2^{nd} Decile, $D_2 = 2 * (n + 1)/ 10^{th}$ data

and so on

Steps to calculate decile:

a. Find out the number of data or variables in the sample or population. This is denoted by ***n***.

b. In the next step, sort all the data or variables in the sample or population in ascending order.

c. In the next step, based on the decile that is required, calculate the decile by using the formula:

$$Di = \frac{i * (n+1)}{10th\ Data}$$

d. Lastly, based on the decile value, determine the corresponding variable from amongst the population data.

Let us look at an example to understand the concept in detail:

Suppose we have been given 23 random numbers between 20 and 80. We need to represent them as deciles.

Let's say the raw numbers are: [24, 32, 27, 32, 23, 62, 45, 77, 60, 63, 36, 54, 57, 36, 72, 55, 51, 32, 56, 33, 42, 55, 30]

Following the steps mentioned above, we first determine the number of variables in the sample (***n***). Here n = 23.

We then need to sort the 23 random numbers in ascending order, as shown below.

SR. No	Digit
1	23
2	24
3	27
4	30
5	32
6	32
7	32
8	33
9	36
10	36
11	42
12	45
13	51
14	54
15	55
16	55
17	56
18	57
19	60
20	62
21	63
22	72
23	77

We can now calculate the positions of decile D_1 to decile D_9.

Now D_1 = 1 * (n+1)/ 10th data

= 1* (23 + 1)/ 10

= 2.4th data i.e. data between digit number 2 & 3

Which is 24 + 0.4 * (27- 24) = 25.2

Similarly,

D_2 = 2 * (n+1)/ 10th data

= 2 * (23 + 1)/ 10

= 4.8th data i.e. data between digit number 4 & 5

Which is 30 + 0.8 * (32 - 30) = 31.6

D_3= 3 * (n+1)/ 10th data

= 3 * (23 + 1)/ 10

= 7.2nd data i.e. data between digit number 7 & 8

Which is 32 + 0.2 * (33 - 32) = 32.2

D_4= 4 * (n+1)/ 10th data

= 4 * (23 + 1)/ 10

= 9.6th data i.e. data between digit number 9 & 10

Which is 36 + 0.6 * (36 - 36) = 36

D_5= 5 * (n+1)/ 10th data

= 5 * (23 + 1)/ 10

= 12th data i.e. data at digit number 12

Which is 45

D_6= 6 * (n+1)/ 10th data

= 6 * (23 + 1)/ 10

= 14.4th data i.e. data between digit number 14 & 15

Which is 54 + 0.4 * (55 - 54) = 54.4

D_7= 7 * (n+1)/ 10th data

= 7 * (23 + 1)/ 10

= 16.8th data i.e. data between digit number 16 & 17

Which is 55 + 0.8 * (56 - 55) = 55.8

D_8= 8 * (n+1)/ 10th data

= 8 * (23 + 1)/ 10

= 19.2nd data i.e. data between digit number 19 & 20

Which is 60 + 0.2 * (62 - 60) = 60.4

D_9= 9 * (n+1)/ 10th data

= 9 * (23 + 1)/ 10

= 21.6th data i.e. data between digit number 21 & 22

Which is 63 + 0.6 * (72 - 63) = 68.4

Thus, we can represent the deciles for the data set with its positions and corresponding values in a table as shown below:

Decile	Data position	Value
1	2.4	25.2
2	4.8	31.6
3	7.2	32.2
4	9.6	36
5	12	45
6	14.4	54.4
7	16.8	55.8
8	19.2	60.4
9	21.6	68.4

One example of the use of deciles is in school rankings. Students in the top 10 % or highest decile will be rewarded, whereas students in the last 10% or lowest decile will be given extra assistance to improve their scores.

Recap

- In Data Science, data merging is the process of combining two or more data sets into a single data frame.
- In one-to-one join, each row in one table is linked to a single row in another table using a "key" column.
- In a one to many join, one record in a table can be related to one or many records in another table.
- A many to many relationships are said to occur when multiple records in one table are related to multiple records of other table.

Exercises

Objective Type Questions

Please choose the correct option in the questions below.

1. The p^{th} percentile of a distribution is such that:
   a) p percent of the observations fall at it
   b) p percent of the observations fall below it
   c) p percent of the observations fall at or below it
   d) the value is p.

2. Which of the following function is used for quantiles of quantitative values?
   a) Quantile
   b) Quantity
   c) Quantiles
   d) All of the mentioned

3. The distribution of heights of Indian women aged 18 to 24 is approximately normally distributed with a mean of 65.5 inches and standard deviation of 2.5 inches. Calculate the z-score for a woman six feet tall.
   a) 2.60
   b) 4.11
   c) 1.04
   d) 1.33

4. What is a z-score?
   a) It is the number of standard deviations a particular score lies above or below the mean of the set of scores.
   b) It is a standardized measure of the mean of a set of data.
   c) It is the average frequency of scores in a sample
   d) It is a measure of central tendency in the data.

5. The median, mode, deciles and percentiles are all considered as measures of
   a) Mathematical averages
   b) Population averages
   c) Sample averages
   d) Averages of position

6. According to percentiles, the median to be measured must lie in
   a) 80^{th}
   b) 40^{th}
   c) 50^{th}
   d) 100^{th}

7. What measures of position divides the distribution into 10 equal parts?
   a) Quartiles
   b) Deciles
   c) Percentiles
   d) Range

8. What measures of position divides the distribution into 4 equal parts?
   a) Quartiles
   b) Deciles
   c) Percentiles
   d) Range

Standard Questions

Please answer the questions below in no less than 100 words.

1. What is data merging?
2. Why is data merging required in data science?
3. Name different ways of merging data sets
4. Explain one-to-one join with the help of an example
5. Explain one-to-many join with the help of an example

6. Explain many-to-many join with the help of an example
7. Think and explain how z-score can be used to determine average lifespan of car tires.

Higher Order Thinking Skills (HOTS)

1. Suppose heights of 2nd graders follow a normal distribution with a mean of 48 inches and a standard deviation of 2 inches. What is the z score of a 2nd grader who is 40 inches tall?

2. Consider that we know about a population of group of plants having their heights that are normally distributed. Further to this, consider that we know that the mean of the distribution is 13cms and the standard deviation is 1.3cms. Now consider the below questions:
    a. What is the z-score for 9cms?
    b. What is the z-score for 2cms?
    c. How many centimeters corresponds to a z-score of 2.25?

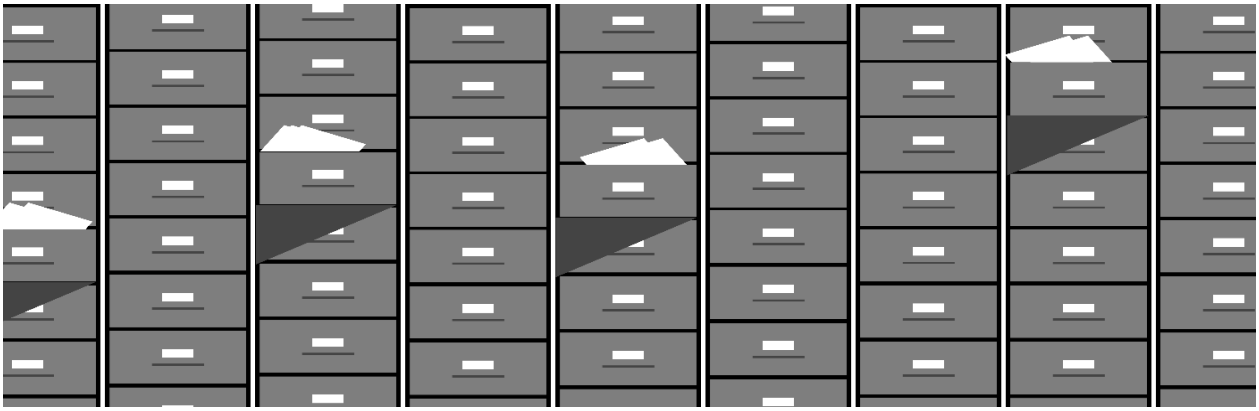
Applied Project

Consider that you are participating in a survey where you are figuring out the probability of a new-born baby to be underweight. Underweight is when a baby is born with a weight less than 2.5kgs. Average weight of a new-born is 3.5kgs and it deviates from average by 1.25kgs. Use z-score to figure out probability of a new-born baby to be underweight.

CHAPTER

5

ETHICS IN DATA SCIENCE



Studying this chapter should enable you to understand:

- Note about data governance framework
- Ethical guidelines around data analysis

1. Note about data governance framework

As we have learnt in the previous grade, data governance framework aims at creating methods, set of responsibilities and processes to standardize, integrate, protect and store data. Moving forward in this chapter, we will learn ethical guidelines around data analysis.

Data analytics raises many ethical issues, especially when anyone starts making money from their data externally for the purposes different from the ones for which the data was initially collected.

Let us now go ahead and understand ethical guidelines around data analysis.

2. Ethical guidelines around data analysis

While there are no specific rules for ethical guidelines around data analysis, below are the principles which experts agree upon:

Protect Your Customer

Privacy does not always mean confidentiality because private data may need to be audited based on the relevant requirements. However, the private data acquired from a person with their

consent should never be exposed for use by different businesses or individuals.

The private information that is shared should always be handled with confidentiality

Third party companies share sensitive data, either financial, location related or medical. They should always have restrictions on if and how that information is allowed to be passed forward.

Customers should always have a clear view of how their data is getting used or traded and should have the authority to manage the flow of their confidential information across enormous, third-party systems.

Data should never interfere with human will

Data analytics can average out and at times, even discover who we are even before we make up our mind. Organizations should begin thinking about the different type of predictions and conclusions that can be allowed and the ones that cannot.

Data should never institutionalize unfair biases like sexism or racism. Analytical systems can absorb unconscious biases in a crowd and boost them with the help of training samples.

3. Discarding the Data

Once we are done with the user data, especially confidential data, it is important that we discard this data in appropriate way to make sure that it is not accessed by any unauthorized person and it is not misused in anyway.

There are two ways in which you may have stored the data – in the digital format or as a physical copy.

It is important to understand that the digital information is the most vulnerable information that you may possess. With the increased amount and intensity of cyber-attacks, it is important for all of us to make sure that we discard the digital data in a proper way. This helps us to prevent unauthorized access to the data.

Once you are done with the job and you no longer need the user data, you can go ahead and clean out the data from the memory.

Even while storing the data in your device, you can encrypt the data to make sure that even in the case of a data leak, hackers are not able to read your data.

You can also format the computer drive/hardisk where the client confidential data was stored for a clean discarding.

Please note that in most of the devices, if you do a soft delete of a particular file, this file deletes from the original space and gets stored in a temporary folder from where one can easily restore these

files. Hence, it is important that confidential data is cleaned out or formatted from the disk permanently and no one is able to restore the files that we have deleted.

Just like digital data, many a times, we may possess the confidential data in the form of physical copies. Now let us understand how can we safely discard the physical copies of confidential data.

Shredding the Documents

Shredding of the documents which contain confidential data is an effective way of discarding the data.

You can use the shredder to shred these documents. It important to make sure that any sensitive information from the shredded documents is unreadable. No person should be able to reconstruct the shredded documents or read the sensitive information from these documents.

Once the documents are shredded, you can be sure that they are effectively discarded.

Cutting up the Documents

In situations where you just have a single page or one file to discard, cutting the documents can be an appropriate method to discard the documents.

While cutting the documents, you should cut them into small pieces and make sure that no sensitive information is readable.

Also, you should cut the document in a way that it is not in a position to be reconstructed and it is completely unreadable.

If these conditions are met, you have successfully discarded the data.

Burning the Documents

Burning the documents is also considered to be an effective way to discard the documents as it makes sure that the documents that are burnt can never be reconstructed or read again.

Although this method is not always practical, it is useful many a times when no other means of discarding are conveniently available.

Recap

- Do not use confidential customer data for business purposes without consent.
- Be transparent with customers on how their data is used.
- Every confidential data that you possess should be appropriately discarded.

Exercises

Objective Type Questions

1) Which of the following is not one of the principles in data governance framework?
   a) Protect your customer
   b) Data should never institutionalize unfair biases
   c) Never collect confidential data from users

2) The private information that is shared should always be handled with confidentiality
   a) True
   b) False

3) If you are done with using the confidential data collected from users, you should :
   a) Safely store it. We may need it in future for some analysis or reports
   b) Effectively destroy it in a way that it is unreadable

4) Confidential data can be stored in which of the following format?
   a) Digital Data
   b) Physical Copies
   c) Both

5) Data should never institutionalize unfair biases
   a) True
   b) False

6) Digital confidential data should be discarded by
   a) Formatting the drive in which data was stored
   b) Temporarily deleting the data

7) Which of the following is not the appropriate way of discarding the confidential data
   a) Shredding the data
   b) Cutting the files which contain confidential data
   c) Burning the confidential data
   d) Crumbling the papers which contain confidential data and throwing it in the dustbin

Standard Questions

1) Explain the significance of data governance framework.
2) Explain principles on ethics that one should follow while performing data analysis.
3) What are various techniques of safely discarding digital confidential data?
4) What are various techniques of safely discarding physical confidential data?

Higher Order Thinking Skills (HOTS)

Please answer the questions below in no less than 200 words.

1) What, according to you, should be the technique used to discard the confidential data collected from users while making an online transaction?
2) How can you make sure that the data that you collected from users while conducting a poll is stored securely?

Applied Project

Suppose you were working with an NGO to help in vaccinating all the kids in your area against Polio. Now that vaccination drive is completed and all the kids are vaccinated, you wanted to make sure that you discard the data about kids that you collected during vaccination. Explain the technique that you would use to discard this data and how will you implement this action.

References

Bargagliotti, A., Franklin, C., Arnold, P. and Gould, R., 2020. Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education II (GAISE II). American Statistical Association

National Research Council. 2013. Frontiers in Massive Data Analysis. Washington, DC: The National Academies Press. https://doi.org/10.17226/18374

National Research Council. 2004. Statistical Analysis of Massive Data Streams: Proceedings of a Workshop. Washington, DC: The National Academies Press. https://doi.org/10.17226/11098

National Academies of Sciences, Engineering, and Medicine. 2020. Collecting and Sharing of Operations and Safety Data. Washington, DC: The National Academies Press. https://doi.org/10.17226/25969

Department of applied mathematics university college, university of london. 2009. Mathematical contributions to the theory of evolution XIII. [Online]. [Accessed 25 February 2021]. Available from: https://archive.org/details/cu31924003064833

Feller, W. (1968). An Introduction to Probability Theory and Its Applications (Third ed.). New York: Wiley. p. 151 (theorem in section VI.3). https://archive.org/details/introductiontopr01wfel

Wadsworth, G. P. (1960). Introduction to Probability and Random Variables. New York: McGraw-Hill. p. 52 https://archive.org/details/introductiontopr0000wads