Content

- **Descriptive Statistics**
    - Measures of Central Tendency
        - Mean
        - Median
        - Mode
    - Measures of Variability
        - Range
        - Variance
        - Standard Deviation
- **Inferential Statistics**
- **Weighted Average**
- **Inter Quartile Range**
    - Quartile
    - Percentile
    - Box Plot
- **IQR implementation on real life dataset**
- **Random Variables**
    - Discrete RV
    - Continuous RV
- **Distribution Functions**
    - Histogram
    - Probability Mass Function (PMF)
    - Probability Density Function (PDF)
    - Cumulative Distribution Function (CDF)

There are 2 types of Statistics:

1. Descriptive Statistics

The word descriptive means "**DESCRIBE**"

Descriptive statistics involve summarizing and presenting data in a meaningful way, providing a clear and concise overview of a dataset.

Example: Let say you are driving a car and you look at your dashboard.

- The speedometer shows the speed of your car at the moment is 65 km/hr. So, it is simply describing speed.
- This Speedometer simply describes an event that a vehicle is moving at a certain speed so it is an example of descriptive statistics.

2. Inferential Statistics

Inferential statistics, on the other hand, involve making predictions, inferences, or drawing conclusions about a larger population based on a sample of data.

"describe"
65 Kmph
"Inference"
You will reach in one hour
120
100
140
80
160
60
180
40
200
20
220
km/h
123456

Continuing the example:

- The car's speedometer displays the current speed but **doesn't predict your arrival time** because it depends on various factors like distance and traffic.
- What Google Maps will do here, is estimate arrival time based on data and assumptions, but it's only sometimes 100% accurate.
- This prediction is an example of inferential statistics, as it draws conclusions from real-world scenarios
- It is trying to "**infer**" something. It is concluding out of it. So, it is inferential statistics.

Conclusion:

- Descriptive statistics summarizes data
- Inferential statistics draws conclusions based on the observations.

These two are the essential branches of statistics.

Let's explore descriptive statistics

Measures of Central Tendency

In statistics, we often use measures to understand and describe a set of data.

Three common measures are:

1. **Mean**
   - The Mean is the average of all data points
2. **Median**
   - The Median is the middle value when the data is sorted.
3. **Mode**
   - It is the observation with the highest frequency

Example-1: Data Scientist's Salaries

```
Suppose we are looking for a data scientist job at FAANG.
The sample of salaries is taken and recorded as [30L, 30L, 35L, 40L, 40L].
What will be the salary we would be expecting?
```

Approach:

1) Mean

- Mean will be (30 + 30 + 35 + 40 + 40)/5 = **35 lakhs**
  - $\mu = \frac{\sum X}{N}$

  Where,

  - μ = population mean
  - $\sum X$ = sum of each value in the population
  - N = number of values in the population

So, the mean salary in this sample is 35 lakhs. We might negotiate our expected salary around this figure.

Suppose, a **new candidate comes** in the context and **his salary is 3 crores**.

- **New mean** will become = (30 + 30 + 35 + 40 + 40 + 300)/6 = **79 lakhs**
- The mean salary dramatically increased to 79 lakhs because of the new candidate's exceptionally high salary.

We can observe that this **new candidate is an outlier** in the data which is affecting the mean value.

2) Median

Here comes the concept of "**Median**" to measure central tendency instead of measuring it using **Mean**.

Before the new candidate joined, the Median was 35L. This means that 35L was the center value when the salaries were sorted in ascending order:

- **Original Salaries:** [30L, 30L, 35L, 40L, 40L]
- **Sorted:** [30L, 30L, 35L, 40L, 40L]
    - **Median** = $35L$ (the middle value)

After the new candidate with a significantly higher salary arrived (300L), the new Median became 37.5 lakhs:

- **New Salaries**: [30L, 30L, 35L, 40L, 40L, 300L]
- **Sorted**: [30L, 30L, 35L, 40L, 40L, 300L]
    - **New Median** = (35L + 40L) / 2 = 75L / 2 = $37.5L$.

Formula:

The median would be:

$(\frac{n+1}{2})th$ observation's value

- For the **even number of observations** the median would be:

$$\frac{(\frac{n}{2})th\ observation + (\frac{n}{2}+1)th\ observation}{2}$$

So, it would be suitable to negotiate at 37.5 lakhs.

There is a **huge difference in the new mean and new median**.

Conclusion:

- The outliers dramatically affect the Mean but the Median remains more robust and closer to the typical value of the dataset.
- Which concludes that **Median is more robust to outliers**

3) Mode

It is the observation with the highest frequency. It is **most occurring data point** in the dataset.

Suppose the data points are recorded as - [90, 90, 90, 80, 90, 70, 95, 90]

- The mode will be **90**.
- Remember, sometimes if there are no data points that repeat, then we can implies that there is no mode

There can also be **more than one mode** in the dataset.

Suppose the data points are recorded as - [2, 2, 3, 3, 4]

- We can call this **Bi-modal** with 2 and 3 as the modes

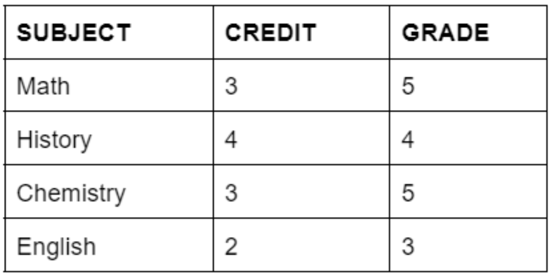
Weighted Average: Reflecting Importance

- In Weighted Average, each data point is assigned a weight that represents its **importance** or relevance.
- We **multiply each data point by its corresponding weight**, **sum these products**, and **then divide by the total weight**.

Example: Calculating GPA

In real life, a common application of weighted average is calculating Grade Point Average (GPA) for students.

Consider a student's course list for a semester:



SUBJECT	CREDIT	GRADE
Math	3	5
History	4	4
Chemistry	3	5
English	2	3

To calculate the GPA:

1. Calculate the weighted score for each course by multiplying the credit by the numerical grade.
2. Sum up all the weighted scores.
3. Divide the total weighted score by the total credits.

Weighted Average will be:

- **For Math:** $3(CREDIT) * 5(GRADE) = 15$
- **For History:** $4 * 4 = 16$
- **For Chemistry:** $3 * 5 = 15$
- **For English:** $2 * 3 = 6$

$$GPA = \frac{Total\ Weighted\ Score}{Total\ Credits}$$

$$= \frac{52}{17} = 3.05$$

Conclusion:

So, the student's GPA for this semester is 3.05

Measures of Variability

Three common measures of variability are

1. Range
2. Variance
3. Standard Deviation

Let's discuss Range

Range

Range is nothing but **Maximum value - Minimum value**

Suppose the Salaries of some employees in a company are : [30, 30, 35, 40, 40]

- Here, the range of the salary will be **40-30 = 10**

It describes the overall spread of the data that the difference between maximum and minimum values is 10.

> **Q1. What will happen if there is an "Outlier" in the data?**

- Let the salaries be: [30, 30, 35, 40, 40, 300]

  New range will be **300 - 30 = 270**

- As we can see one outlier can destroy the range of the dataset.

We can conclude that **Range of the data is also not robust to the outliers like the Mean**

To solve this issue, statisticians came up with the metric called "**Inter Quartile Range**".

Inter Quartile Range

IQR is the metric that provides a robust way to measure the spread of a dataset.

- The IQR is the range between the first quartile (Q1) and the third quartile (Q3) of a dataset.
- Means, $IQR = Q3 - Q1$

> What is Quartiles?

Quartiles

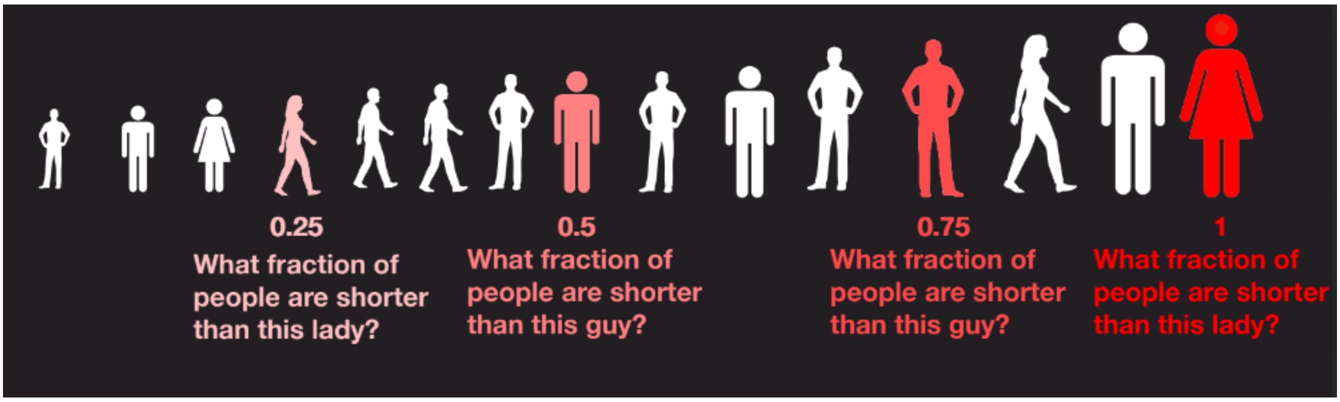
It is the value which divides the dataset into four equal parts.

There are three quartiles, **Q1, Q2, and Q3**.

- Q1 represents the 25th percentile, meaning that 25% of the data falls below this value.
- Q2 is the median and represents the 50th percentile, dividing the data into two equal halves.
- Q3 represents the 75th percentile, meaning that 75% of the data falls below this value.

Suppose we have this data with us,

> **What each values are representing here?**


0.25
What fraction of people are shorter than this lady?
0.5
What fraction of people are shorter than this guy?
0.75
What fraction of people are shorter than this guy?
1
What fraction of people are shorter than this lady?

- 0.25 represents Q1 or 25th percentile:
  - It means that 0.25% of people are shorter than this lady
- 0.5 represents Q2 or 50th percentile:
  - It means that 0.5% or half of the people in the dataset are shorter than this guy
- 0.75 represents Q3 or 75th Percentile:
  - It means that 0.75% of people are shorter than this guy
- Maximum or 1:
  - It means that 100% of people are shorter than this lady or she is the tallest person in the dataset.

> Q.What is percentile?

Percentile

A value that tells us that some **"p%" observations are less than that value**

- Let's say the value occurring at 50 Percentile is 68. We can conclude that 50% of the data is less than 68

One more example:

> Suppose you scored 99 percentile in your 10th boards, what does this mean?

- It indicates that **99% of students scored less marks than you**.

The graphical representation of a dataset's summary statistics, including the median, quartiles, and potential outliers. Box plot comes into the picture

Box Plot

It provides a visual way to understand the distribution and spread of data.

Box:

- The box itself represents the interquartile range (IQR),
  It is divided into two parts, the lower (bottom) quartile (Q1) and the upper (top) quartile (Q3).
- The length of the box is determined by the range between Q1 and Q3.

Line (Median):

- Inside the box, a line or bar is drawn that represents the median, which is the middle value of the dataset when it's ordered.

Whiskers:

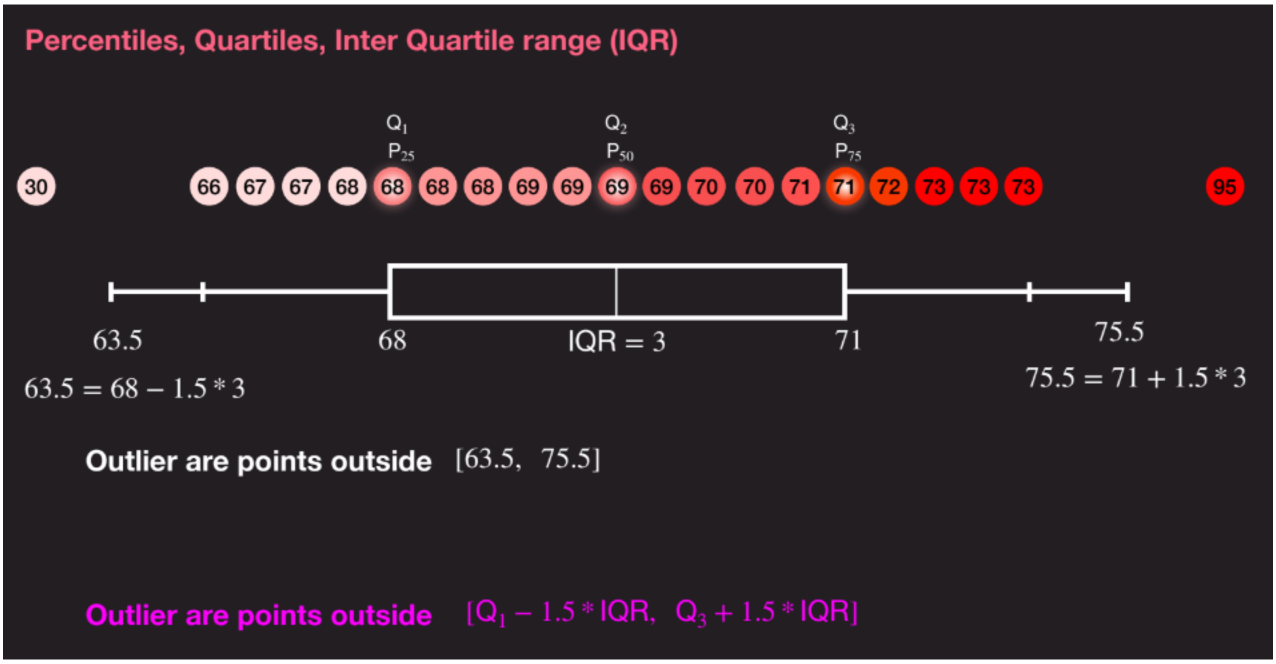
- Two lines, or "whiskers," extend from the box in both directions.
  - The Value of Lower whiskers is determined by Q1 - 1.5(IQR). It is the minimum value of the range
  - The Value of Upper whiskers is determined by Q3 + 1.5(IQR). It is the maximum value of the range

Outliers:

- Data points that fall outside the whiskers are considered outliers.

Example

We have a sorted data, the box plot of the data will look like this



Percentiles, Quartiles, Inter Quartile range (IQR)

$63.5 = 68 - 1.5 * 3$

$75.5 = 71 + 1.5 * 3$

Outlier are points outside $[63.5,\ \ 75.5]$

Outlier are points outside $[Q_1 - 1.5 * \text{IQR},\ \ Q_3 + 1.5 * \text{IQR}]$

IQR implementation on a real-life dataset

Problem Statement:

When we talk about these two players:

1. Sehwag
2. Rahul Dravid

We all know that Sehwag has **aggressive batting style**

While Rahul Dravid **plays patiently**, with no risk and stands on the crease like a "Wall"

- Let's analyse both of their matches and try to find some insights about their range of scores.
- We will use IQR here to calculate the range of their scores accurately and will also try to find if they have any "Outlier" scores in their careers.

We will conclude that out of these two batsman, who is the more consistent batsman?.

Let's start with Sehwag's matches

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
!wget https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/035/130/original/sehwag.csv?1684996594 -O sehwag.csv
```

```
--2023-10-31 06:00:12--  https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/035/130/original/sehwag.csv?168
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)... 18.172.139.94, 18.172.139.46, 18.172.139.210,
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)|18.172.139.94|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 18584 (18K) [text/plain]
Saving to: ‘sehwag.csv’

sehwag.csv          100%[===================>]  18.15K  --.-KB/s    in 0.002s

2023-10-31 06:00:13 (11.0 MB/s) - ‘sehwag.csv’ saved [18584/18584]
```

```
sehwag = pd.read_csv("sehwag.csv")
sehwag.head()
```

	Runs	Mins	BF	4s	6s	SR	Pos	Dismissal	Inns	Unnamed: 9	Opposition	Ground	Start Date	Unnamed: 13
0	1	5	2	0	0	50.00	7	lbw	1	NaN	v Pakistan	Mohali	1 Apr 1999	ODI # 1427
1	19	18	24	0	1	79.16	6	caught	1	NaN	v Zimbabwe	Rajkot	14 Dec 2000	ODI # 1660
2	58	62	54	8	0	107.40	6	bowled	1	NaN	v Australia	Bengaluru	25 Mar 2001	ODI # 1696
3	2	7	7	0	0	28.57	6	caught	2	NaN	v Zimbabwe	Bulawayo	27 Jun 2001	ODI # 1730
4	11	19	16	1	0	68.75	6	not out	2	NaN	v West Indies	Bulawayo	30 Jun 2001	ODI # 1731

```
sehwag["Runs"].describe()
```

```
count    245.000000
mean      33.767347
std       34.809419
min        0.000000
25%        8.000000
50%       23.000000
75%       46.000000
max      219.000000
Name: Runs, dtype: float64
```

We want to find the range of his scores

> Let's find Quartiles first on the "Runs" column.

So Q1, Q2 and Q3 will be

```
# 25th percentile or Q1
p_25 = np.percentile(sehwag["Runs"], 25)
p_25
```

```
8.0
```

This value indicates that 25% of all the values present in the dataset for Sehwag's run is less than 8

We can also say,

Out of all the matches that Shewag played, in 25% of those matches, he scored less than 8 runs.

```
#50th percentile or Q2, also "Median"
p_50 = np.percentile(sehwag["Runs"], 50)
p_50
```

```
23.0
```

This indicates that in 50% of the matches, he scored less than 23 runs

```
#75th percentile or Q3
p_75 = np.percentile(sehwag["Runs"], 75)
p_75
```

```
46.0
```

This indicates that in 75% of the matches, he scored less than 46 runs

> So, IQR will be?

We know IQR = Q3 - Q1

```
# Inter Quartile Range
iqr_sehwag = p_75 - p_25
iqr_sehwag
```

```
38.0
```

```
normal_range = (sehwag["Runs"].max() - sehwag["Runs"].min())

normal_range
```

```
219
```

We can observe the difference here,

IQR is 38 which means that middle 50% of the data lies in the range of 38.

So more than 50% of the time, Sehwag scores in the range of 38 runs

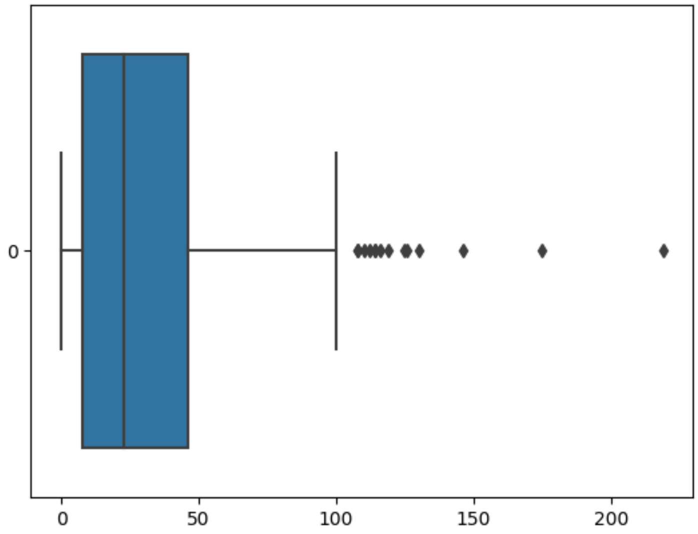
- On the other hand, the **normal range is very high i.e. 219** which is certainly not a good range to consider.
- We can observe one thing that **there in an Outlier present in the data** means in some matches he has scored so many runs like more than 300 in a single match

This is why the range is getting affected by the outlier

Let's plot the box plot to visualise the spread of the data

```
sns.boxplot(data=sehwag["Runs"], orient="h")
```

```
<Axes: >
```



We can see that Q1, Q2, and Q3 values lie within the box and we can also see whiskers on both the sides of box which is the limit.

We already saw how to calculate the lower whisker and upper whisker

All the values outside the limit are considered "Outlier"

```
# upper limit = Q3 + 1.5 * IQR

upper = 46 + 1.5*(iqr_sehwag)
upper
```

```
103.0
```

Here, we cannot have values on the left side of the lower whisker as the batsman cannot score less than 0 runs.

So all the outliers will be present on the right side of the upper whisker

```
# all the values greater than upper is outlier
outliers_sehwag = sehwag[sehwag["Runs"]>upper]
len(outliers_sehwag)
```

```
14
```

```
14/245
```

```
0.05714285714285714
```

Conclusion:

Here we can observe that **5.7% values from the dataset are outliers**.

This means we can conclude that 5.7 or ~6% times Sehwag has scored more than the IQR which is 38 runs

Now let's have a same process into Dravid's stats

Cricket - Dravid

```
!wget https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/035/131/original/dravid.csv?1684996749 -O dravid.csv
```

```
--2023-10-31 06:00:13--  https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/035/131/original/dravid.csv?168
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)... 18.172.139.94, 18.172.139.46, 18.172.139.210,
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)|18.172.139.94|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 24177 (24K) [text/plain]
Saving to: ‘dravid.csv’

dravid.csv          100%[===================>]  23.61K  --.-KB/s    in 0.001s

2023-10-31 06:00:13 (32.8 MB/s) - ‘dravid.csv’ saved [24177/24177]
```

```
dravid = pd.read_csv("dravid.csv")
```

```
dravid["Runs"].describe()
```

```
count    318.000000
mean      34.242138
std       29.681822
min        0.000000
25%       10.000000
50%       26.000000
75%       54.000000
max      153.000000
Name: Runs, dtype: float64
```

```
#25th percentile or Q1
per_25 = np.percentile(dravid["Runs"], 25)
per_25
```

```
10.0
```

This indicates that in 25% of the matches, he scored less than 10 runs

```
#50th percentile or Q2 , also "Median"
per_50 = np.percentile(dravid["Runs"], 50)
per_50
```

```
26.0
```

This indicates that in 50% of the matches, he scored less than 26 runs

```
#75th percentile or Q3
per_75 = np.percentile(dravid["Runs"], 75)
per_75
```

```
54.0
```

This indicates that in 75% of the matches, he scored less than 54 runs

```
# Inter Quartile Range
iqr_dravid = per_75 - per_25
iqr_dravid
```

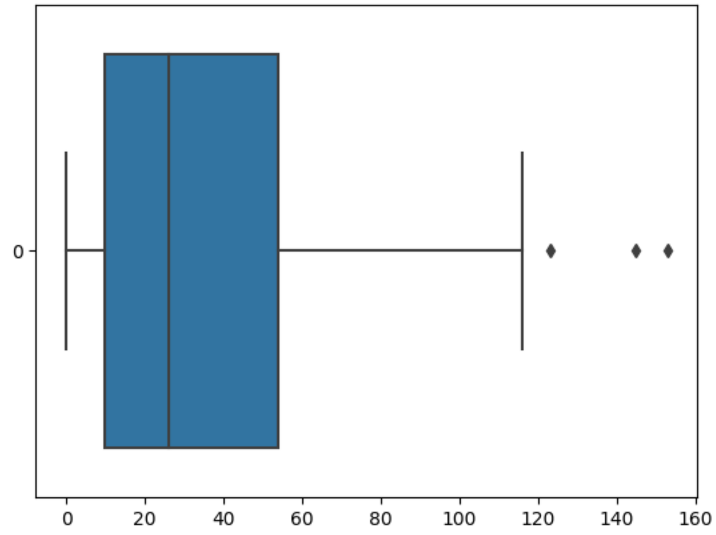
```
44.0
```

```
normal_range = (dravid["Runs"].max() - dravid["Runs"].min())

normal_range
```

```
153
```

```
 sns.boxplot(data=dravid["Runs"], orient="h")
```

```
<Axes: >
```



```
# upper limit = Q3 + 1.5 * IQR

upper_dravid = per_75 + 1.5*(iqr_dravid)
upper_dravid
```

```
120.0
```

```
# all the values greater than upper is outlier
outliers_dravid = dravid[dravid["Runs"]>upper_dravid]
len(outliers_dravid)/len(dravid)
```

```
0.009433962264150943
```

```
outliers_dravid['Runs'].shape
```

```
(3,)
```

```
dravid.shape
```

```
(318, 14)
```

Conclusion

Here we can observe that **0.9% values from the dataset are outliers**.

This means we can conclude that 0.9% times Dravid has scored more than the IQR which is 44 runs

So we can conclude that in Sehwag case there is **6% outliers** and in Dravid's case there are only **0.9% outliers** which shows that "**Dravid was more consistent than Sehwag**"

Random Variable (RV):

A random variable is a situation/event/experiment, for which we are not certain about the outcome.

It is a way to assign numbers to the outcomes of such events.

They can further be divided into 2 types:

- Discrete Random Variable
- Continuous Random Variable

Examples of Discrete Random Variable

Here, we can count the number of possible outcomes.

> **1. Coin Toss**

Let's consider a coin toss.

What are its possible outcomes?

Heads and Tails.

- There is no other possible other than this.

Hence, we can represent its outcomes as a random variable, that can take values: $\{H, T\}$

> **2. Throw of a dice**

- Let's assign a random variable, "X," to represent the outcome of the die roll.
- So, a throw of dice can be represented as: $X = \{1, 2, 3, 4, 5, 6\}$, depending on the outcome of the roll.
- It can not have an outcome lesser than 1, or greater than 6
- Or even, any decimal value between 1 and 2
- Hence it is also discrete RV

Examples of Continuous Random Variable

Here, we cannot count the number of possible outcomes. They are infinite.

> **1. Height of students in a class**

- Suppose the lowest student height in the class is: 4.5 feet
- Suppose the highest student height is: 5.9 feet

Now, we can have students that have height as

- 4.511 feet
- 4.92 feet
- 5.8555 feet

So, we have an infinite number of possible height values between 4.5 and 5.9 feet. We cannot count them

Whereas, we could count the number of possibilities in a coin toss or dice throw.

> **Other examples of Continuous RV can be:**
>
> - Temperature of a room
> - Time taken to complete a task
> - Distance travelled

...etc

Distribution Functions

Probability Density Function (PDF):

- The PDF is a function that describes the probability density of a continuous random variable over its range.
- The term "density" here is similar to how tightly data is packed around a specific point, like cars on a road.

Probability Mass Function (PMF):

- The PMF is a function that describes the probability of a discrete random variable taking on a specific value.

Cumulative Distribution Function (CDF):

- The CDF is a function that gives the probability that a random variable is less than or equal to a specified value.

Let's implement this using a height dataset

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
!wget https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/035/126/original/weight-height.csv?1684995383 -O weigh
```

```
--2023-10-31 06:00:13--  https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/035/126/original/weight-height.c
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)... 18.172.139.94, 18.172.139.46, 18.172.139.210,
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)|18.172.139.94|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 428120 (418K) [text/plain]
Saving to: ‘weight-height.csv’

weight-height.csv   100%[===================>] 418.09K  --.-KB/s    in 0.06s

2023-10-31 06:00:13 (6.84 MB/s) - ‘weight-height.csv’ saved [428120/428120]
```

```
df_hw = pd.read_csv("weight-height.csv")
```

```
df_hw.head()
```

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801

```
df_hw.describe()
```

	Height	Weight
count	10000.000000	10000.000000
mean	66.367560	161.440357
std	3.847528	32.108439
min	54.263133	64.700127
25%	63.505620	135.818051
50%	66.318070	161.212928
75%	69.174262	187.169525
max	78.998742	269.989699

We will going to work on the single column for now

```
df_height = df_hw["Height"]
df_height.head()
```

```
0    73.847017
1    68.781904
2    74.110105
3    71.730978
4    69.881796
Name: Height, dtype: float64
```

```
# minimum height
min_height = df_height.min()
min_height
```

```
54.2631333250971
```

```
# maximum height
max_height = df_height.max()
max_height
```

```
78.9987423463896
```

```
total = len(df_height)
total
```

```
10000
```

When we talk about probability, we try to construct the Distribution plots.

Cumulative Distribution Function (CDF), Probability Mass Function (PMF), and Probability Density Function (PDF) are all related to random variables and are used to describe the probability distribution of random variables.

First, let's see what is random variable

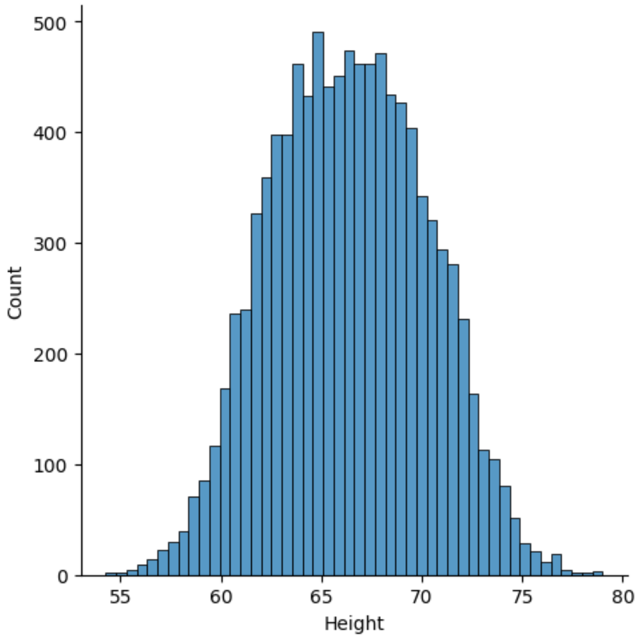
To plot this type of distribution we generally use Histograms or Distribution plots

Histogram

It is a graphical representation of a dataset's distribution, showing the frequency or probability of different values within the data.

```
sns.displot(df_height)
```

```
<seaborn.axisgrid.FacetGrid at 0x7c3b394ab130>
```



> Q.What we can understand from this distribution?

- Each bar in the histogram represents one of the intervals or ranges,
- The height of the bar indicates the frequency or number of data points falling within that interval.

Count:

- It indicates the "**frequency**", which means in the particular bar or range of height, how many values are there.
  - We can observe that, around 500 people have their height in the range of 63 - 65 (that on bar)

This is what histograms or distribution plots tell about the data

1. Probability Mass Function (PMF)

The PMF is a function that describes the probability of a discrete random variable taking on a specific value.

It associates each possible value of the random variable with its probability of occurrence.

Example: Rolling a Fair Six-Sided Die

- If we have a discrete random variable X representing the outcome of rolling a fair six-sided die

Possible outcome is: 1, 2, 3, 4, 5, 6. This is discrete random variable

- The PMF might look like $P(X = 1) = \frac{1}{6}, P(X = 2) = \frac{1}{6}$, and so on.

2. Probability Density Function (PDF)

PDF is used for continuous random variables, as opposed to PMF, which is for discrete variables.

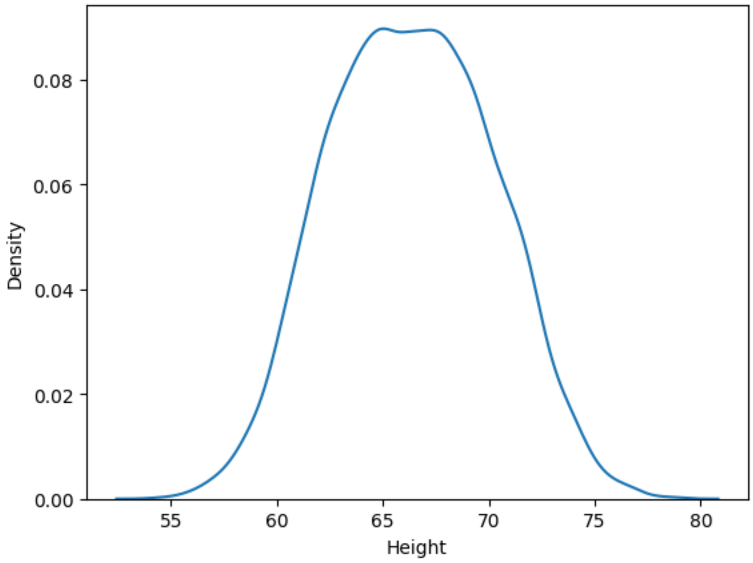
If we want to find the probability of a specific value which is continuous random variable within the given range then we will use PDF.

- It doesn't provide the probability of a specific value but gives us the probability of the RV falling within a certain interval.
- For example, what are the chances that the next height we chose will fall between 62 and 65

We can visualize a PDF by using distribution plots like histograms or KDE (Kernel Density Estimation) plots.

```
sns.kdeplot(df_height)
```

```
<Axes: xlabel='Height', ylabel='Density'>
```



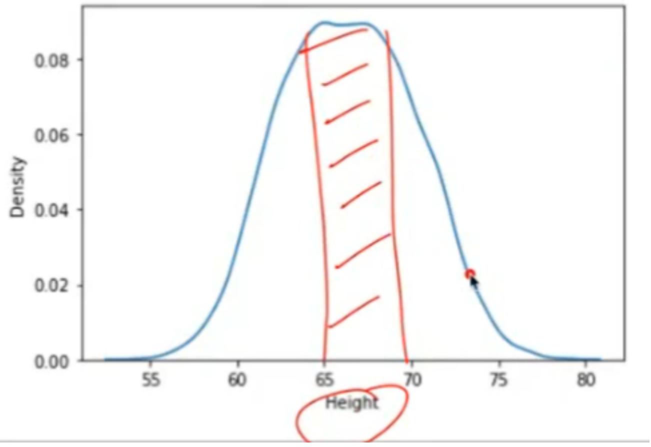
Example:

If we have a continuous random variable Y representing the height of people in a population,

The PDF might represent the probability that a randomly chosen person has a height within a certain range, such as between 65 and 70.

- We will find out the area under that interval to find the probability

```
<AxesSubplot:xlabel='Height', ylabel='Density'>
```



> **Q. Can we find the probability of exact values using PDF?**

- As we know Probability Distribution Function (PDF) works with continuous random variables
- It represents the likelihood of a random variable falling within a particular interval, not at a precise point.
- So, the PDF doesn't assign probabilities to exact values since there are infinitely many possible values within a continuous range.

In other words, we compute the probability that X falls in a range [a, b].

In summary, the PDF represents probabilities as density over an interval rather than exact values

3. Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) describes the probability that a random variable takes on a value less than or equal to a given value.

In the context of this dataset, in CDF, we talk about fractions of people who are less than the given height

- Let's say we take 60 inches, then what fraction of the people have less than or equal to this value? This fraction is calculated using CDF
- It gives us the cumulative probability up to a certain point.

Example:

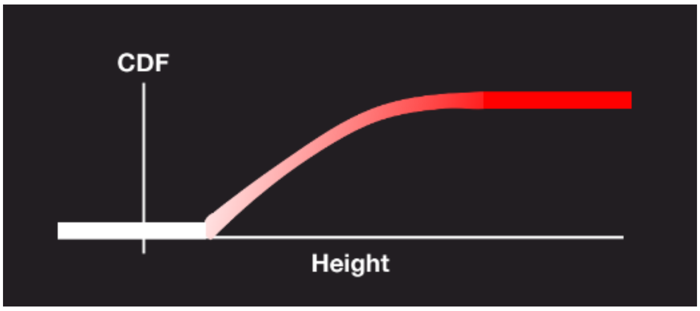
If we have a random variable Z representing the number of heads in three coin tosses,

The CDF would tell us the probability that Z is less than or equal to a certain number, like P(Z ≤ 2).

> **How to calculate CDF**?

The **CDF is calculated by accumulating the probabilities for each height value**.

- As we move along the X-axis (height values) on the CDF graph, we're essentially adding up the probabilities
- It shows how likely it is to find someone with a height less than or equal to that value.



- The CDF graph typically starts at 0% on the Y-axis (probability) when height is at its minimum (in our dataset)
- It ends at 100% when height is at its maximum.

- The curve starts at the left and gradually climbs towards the right.
- The steepness of the curve at a particular point represents how quickly the probability is accumulating

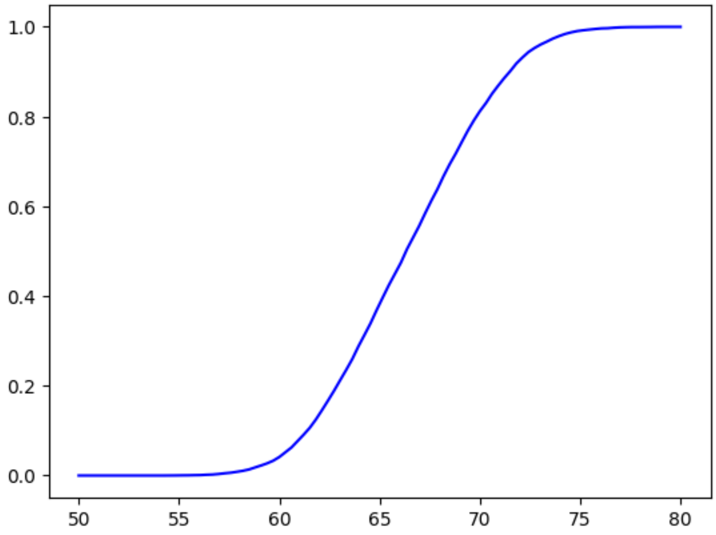
Conclusion

So, the PDF shows us the probability of a specific height, while the CDF shows us the probability of heights up to a certain value in our dataset.

Let's plot the CDF graph for this dataset manually

```
# CDF: Cumulative distribution function

# will take 100 values between the range of 50 and 80 inclusively using np.linpace
x_values = np.linspace(50, 80, 100)

# Will contain fraction of people shorter than x
y_values = []

for x in x_values:

    # find out people shorter than x
    people_shorter_than_x = df_height[df_height <= x]

    # find out number of such people
    num_people_shorter_than_x = len(people_shorter_than_x)

    # How many fraction of people are shorter than x so dividing it by total value
    fraction_people_shorter_than_x = num_people_shorter_than_x / total

    # Appending into the y_values list
    y_values.append(fraction_people_shorter_than_x)

# plotting the CDF
plt.plot(x_values, y_values, c="b")
```

```
[<matplotlib.lines.Line2D at 0x7c3b3c1d0490>]
```



- This Curve is called "**Cumulative Distribution Function**".
- It is a function which takes the x value and returns the y value
    - $f(x) = y$

It is the inverse of the percentile means

- **Percentile will take input as 25 and give output as 63.5** means
    - 25% of the people are shorter than 63.5
- While **CDF will take input as 63.5 and give output as 0.25** means
    - if we want to find how many people are having height less than or equal to 63.5 i.e. 25% of people.

Conclusion :

In summary, the relationships are as follows:

- The PMF is used for discrete random variables.
- The PDF is used for continuous random variables.
- The CDF is used for both discrete and continuous random variables to provide cumulative probabilities.

These functions are essential tools in probability and statistics for describing and understanding the behaviour of random variables.

Variance

Variance, measures the spread or dispersion of the values of a random variable around its mean.

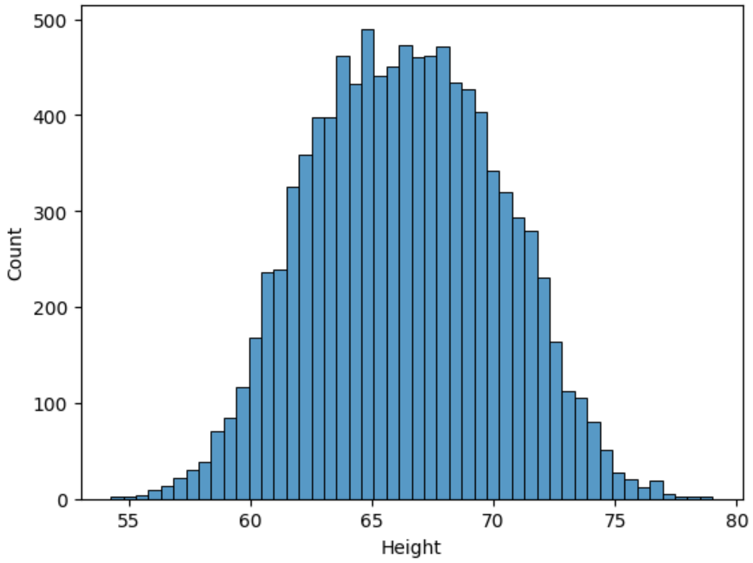
It quantifies how much **individual values deviate from the mean**.

- A **higher variance** indicates that the values are more spread out from the mean.
- While a **lower variance** suggests that the values are closer to the mean.

We can plot the histogram to visualise the spread or distribution of the data

```
sns.histplot(df_height)
```

```
<Axes: xlabel='Height', ylabel='Count'>
```



Let's explore another way to measure error

Defining Error:

$Error = (Actual\ Height - Guessed\ Height)^2$

> **Q1. Now, to minimize this error, what's the best approach?**

In our Height Guessing Game, we've seen that aiming for the mean (μ) height is the key. Means, Guessed height should be the mean value.

- $Error = (H_1 - \mu)^2$ (guessing for 1 time)
- It is also known as **Mean Squared Error**

Imagine we're playing the game 10 times, guessing the mean height each time:

Error1 = $(H1 - \mu)^2$

Error2 = $(H2 - \mu)^2$

Error3 = $(H3 - \mu)^2$

...

Error10 = $(H10 - \mu)^2$

To find the overall error, we can sum up these individual errors and then divide by the number of guesses, which gives us the variance:

Variance Calculation:

Variance = (Error1 + Error2 + Error3 + ... + Error10) / 10

- $V ariance = \frac{(H_1-\mu)^2+(H_2-\mu)^2+(H_3-\mu)^2+.....+(H_{10}-\mu)}{10}^{2}$

> **Q2. So, if the variance is low, what does that mean?**

It implies that most of our guesses are incredibly accurate.

In general, variance quantifies how spread out, the data values are from the average (mean) value.

It assesses the average squared difference between data points and the mean.

The formula for calculating variance for n data points is:

Variance Calculation Formula:

$$variance = \sigma^2 = \frac{\sum_{i=1}^{n}(H_i - \mu)^2}{n}$$

- σ2 is the population variance.
- H_i is the ith data point.
- μ is the population mean.
- n is the number of data points in the population.

Now that we have a clear understanding of variance and how it measures the spread or dispersion of data points,

Let's look into another essential concept closely related to variance. It's called "**Standard Deviation**"

Standard Deviation

Let's introduce an even more practical and commonly used statistic - the "Standard Deviation."

While variance quantifies the dispersion of data, standard deviation is derived from variance and offers a more interpretable measure.

- The standard deviation represents how much individual data points deviate from the mean or average value.
- It gives us a clear sense of the typical or expected amount of variation in our dataset.
  - In simple words, it represents that how far is our data point from the mean (μ)

Standard Deviation Formula:

The standard deviation, can be calculated by taking the square root of the variance:

$SD = \sqrt{variance}$

- $$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(H_i - \mu)^2}{n}}$$

Interpretation:

- A lower standard deviation signifies that data points tend to be close to the mean, indicating **less variability**.
- Conversely, a higher standard deviation indicates greater data dispersion, suggesting **more variability** within the dataset.