Descriptive Statistics- Measure of Central Tendency and Measures of Dispersion

Relevel
by Unacademy

Different type of Statistics

In this section, we will cover overview on :

- Descriptive Statistics
- Inferential Statistics
- Deep-Dive on Descriptive Statistics

Categories of Statistics

Population: All the events in an experiment is considered population.
Sample: Pulling out some events out of it

Statistics is majorly categorised into two types:

1. Descriptive statistics: Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures.

For example;- reporting a general trend about the collection of people in a city using the internet or television. This reporting would come under descriptive analytics.

2. Inferential statistics: With the help of inferential statistics, we can use data from a sample to extrapolate conclusions about the population. It enables us to make claims beyond the scope of the facts or data at hand.

For example, deriving inferences from our report on the trends in people using the internet or TV, we can infer that people below 40 years of age use the internet more than TV. In contrast, people above 40 years of age use TV more.

** **In this session, we will deep-dive into descriptive statistics.***

Overview of Descriptive Statistics

Descriptive statistics is also categorised into four different categories:

- **Measure of Frequency:** The frequency measurement displays the number of times a particular data occurs(histogram, frequency distribution). This is already covered in the previous class.

- **Measure of Central tendency:** It gives the central value of the distribution.

- **Measure of Position:** It tells the position of a single value about other values in a sample or a population dataset.

- **Measure of Dispersion:** Dispersion represents the spread of data, basically what range of values the variable can take.

Measure of Central Tendency

In this section, we will cover ::

- Mean
- Median
- Mode
- Practice Problem

Measure of Central Tendency

Measures of central tendency describe a data set by identifying the central position in the data set as a single representative value. There are generally three measures of central tendency, commonly used in statistics

1. Mean
2. Median
3. Mode

We come across new data every day. We find them in newspapers, articles, bank statements, and mobile and electricity bills. Now the question arises whether we can figure out some important features of the data by considering only certain data representatives. This is possible by using measures of central tendency. In the following sections, we will look at the different measures of central tendency and the methods to calculate them.

Mean

Mean is the sum of all the components in a group or collection, divided by the number of components. It is also known as average.

The formula to calculate the mean for ungrouped data to represent it as the measure is given as,
For a set of observations: **Mean = Sum of the terms/Number of terms**
For a set of grouped data: **Mean, $\bar{x}$ = $\Sigma fx / \Sigma f$**
where,

- $\bar{x}$ = the mean Value of the set of given data.
- f = frequency of each observation
- x = Value of each observation (in case we go for a frequency bin distribution like we did for the continuous data in the last class, x will take the mid-interval value)

Don't get worried by the Σ; it simply means sum!

Example of Mean

Example: Calculate the mean in the following example.
In a quiz, the marks obtained by 20 students are given as:
12,15,15,29,30,21,30,30,15,17,19,15,20,20,16,21,23,24,23,21
Average = sum of all observations / total number of observations
Here, Sum of all observations = 12+15+15+29+30+21+30+30+15+17+19+15+20+20+16+21+23+24+23+21
= 416
For later, can we also write this as - 12 + 15 + 15 + 15 + 15 + 16 + 17+.... by rearranging the data?
This is equal to - 12 + (15*4) + 16 + 17 + 19 + (20*2) + (21*3) + (23*2) + 24 + 29 + (30*3) = 416

Number of observations = 20
Mean = 416/20 = 20.8
Now let's understand the formula
Mean, $\bar{x} = \Sigma fx / \Sigma f$
For this, we first express the data in a frequency distribution table

Calculate the mean of the following data

In a quiz, the marks obtained by 20 students are given as:
12,15,15,29,30,21,30,30,15,17,19,15,20,20,16,21,23,24,23,21

X-values	Marks obtained in quiz	Number of students (Frequency)	F values
X1	12	1	F1
X2	15	4	F2
X3	16	1	F3
X4	17	1	F4
X5	19	1	F5
X6	20	2	F6
X7	21	3	F7
X8	23	2	F8
X9	24	1	F9
X10	29	1	F10
X11	30	3	F11
	Total	20	ΣF

For a set of grouped data: Mean, $\bar{x} = \Sigma fx/\Sigma f$
where,
$\bar{x}$ = the mean value of the set of given data.
f = frequency of each observation
x = Value of each observation

Here, this will be
$\bar{x}= (f1*x1 + f2*x2 + f3*x3)+...+ f11*x11)/\Sigma f$
$= (12*1 + 15*4 + 16*1 + 17*1 +...+ 30*3)/20$

Here, Σfx is the same as we made in the previous slide!
$\Sigma fx = 416$
$\Sigma f = 20$
$\bar{x} = 20.8$

Example of Mean – Grouped Data

Calculate the mean of this frequency table.

X	f
1	15
2	27
3	8
4	5

Solution

$x = \Sigma fx/\Sigma f$

$= (15\times1 + 27\times2 + 8\times3 + 5\times4)/15+27+8+5$

$=2.05$

Median

The value of the provided data-set that is the middle-most observation after the **data are arranged in ascending order** is known as the median, another measure of central tendency.
The median is less impacted by outliers and skewed data, which is a primary benefit of using it as a central tendency. Using the median formula, we can determine the median for various data types, grouped data, and ungrouped data.

Calculating Median

For ungrouped data:

- For odd number of observations,

Median = **[(n + 1)/2]th term**

- For even number of observations,

Median = **[(n/2)th term + ((n/2) + 1)th term]/2**

For grouped data:

Median = **l + [((n/2) - c)/f] × h**

where,

l = Lower limit of the median class

c = Cumulative frequency

h = Class size

n = Number of observations

Median class = Class where n/2 lies

Example of Median – For Ungrouped data (Even)

Example: The weights of 8 boys in kilograms: 46, 39, 53, 45, 43, 48, 50, 47. Find the median.

Solution:

Arranging the given data set in ascending order: 39, 43, 45, 46, 47, 48, 50, 53
Total number of observations = 8

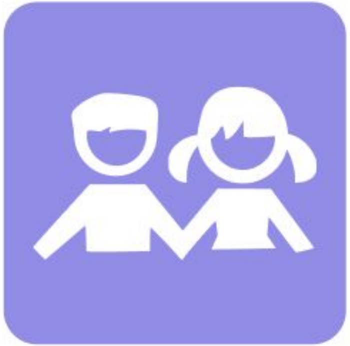
For even number of observation, Median = [(n/2)th term + ((n/2) + 1)th term]/2

Median = (4th term + 5th term)/2 = (45 + 46)/2 = 45.5

Example of Median – For Ungrouped data (Odd)

Example: The age of 9 boys in years: 5, 2, 8, 6, 7, 15, 19, 13,10. Find the median.

Solution:



Arranging the given data set in ascending order: 2, 5, 6, 7, 8, 10, 13, 15, 19
Total number of observations = 9

For odd number of observation, Median = [(n + 1) / 2] th term

Median = [(9+1)/2]th term = 5th term
= 8

Example of Median – For grouped data

The following data represents the survey regarding the heights (in cm) of 51 girls of Class x. Find the median height.

Height (in cm)	Number of Girls
Less than 140	4
Less than 145	11
Less than 150	29
Less than 155	40
Less than 160	46
Less than 165	51

Example of Median – For grouped data (solution)

- We must first identify the class intervals and their related frequencies before getting the median height.
- The supplied distribution is less than typical, with upper limits of 145, 150, and 165. Therefore, the class should be between 140 and 145, 150 and 155, and 160 and 165.
- Four girls fall below the 140-point threshold, according to the supplied distribution. Consequently, there are four class intervals below 140 on average.
- Four girls under 140 inches tall and 11 under 145 inches tall.
- As a result, the frequency distribution for the range of classes 140 to 145 is 11-4 = 7.
- The frequency of 145 - 150 = 29 - 11 = 18 is similar.
- 150-155 = 40-29 = 11 frequency
- 155 - 160 = 46-40 = 6 frequency
- Frequency of 160-165 = 51-46 = 5

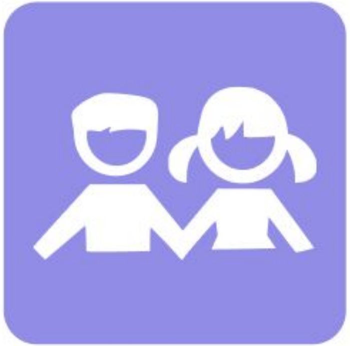
Example of Median – For grouped data (solution)

The frequency distribution table along with the cumulative frequencies are given below:

Class Intervals	Frequency	Cumulative Frequency
Below 140	4	4
140 – 145	7	11
145 – 150	18	29
150 – 155	11	40
155 – 160	6	46
160 – 165	5	51

Example of Median – For grouped data (solution)

Here, n= 51.
Therefore, n/2 = 51/2 = 25.5
Thus, the observations lie between the class interval 145-150, which is called the median class.
Therefore,
Lower class limit = 145
Class size, h = 5
Frequency of the median class, f = 18
Cumulative frequency of the class preceding the median class, cf = 11.
We know that the formula to find the median of the grouped data is:
Now, substituting the values in the formula, we get:
Median = 145 + (72.5/18)
Median = 145 + 4.03
Median = 149.03.
Therefore, the median height for the given data is 149. 03 cm.



Mode

- One of the measures of central tendency is the mode, which is defined as the value that occurs the most frequently in the supplied data or, more specifically, as the observation with the highest frequency. Using the mode formulas provided below, one can determine the mode for grouped or ungrouped data.

- The most frequent observation in the data set serves as the mode for ungrouped data.

- Data grouping mode:

$$L + h \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)}$$

where L is the modal class lower limit.

f_m is the frequency of the modal class, and h is the size of the class interval.

- The frequency of the class that comes before the modal class is f_1, and the frequency of the class that follows the modal class is f_2.

Example of Mode: Ungrouped Data

The weights of 8 boys in kilograms: 45, 39, 53, 45, 43, 48, 50, 45. Find the mode.

Solution:

Since the mode is the most occurring observation in the given set,

Mode = 45

Example of Mode: Grouped Data

Find the mode of the given data:

Class	Frequency
50-55	2
55-60	7
60-65	8
65-70	4

Example of Mode: Grouped Data (Solution)

Here 8 is the maximum class frequency

Here modal class is 60-65

Class size, h = 65-60 = 5

Mode = $l + [(f_m – f_1)/((f_m – f_1)+(f_m – f_2))]h$

$f_m = 8$

$f_1 = 7$

$f_2 = 4$

$l = 60$

$h = 5$

So mode = $60 + ((8-7)/(16 – 7 – 4)]5$

$= 60 + (1/5)5$

$= 61$

Practice Problem -1

Example 1: The average monthly wage for a group of ten employees is \$1445. One more worker whose monthly salary is \$1500 has joined the group. Find the average monthly wage for 11 employees.

Practice Problem -1 (Solution)

Solution 1:

Here, $\sum f=10$, $\overline{x}=1445$

Using the formula,

$\overline{x} = \sum fx/\sum f$

Therefore $\sum fx = \overline{x} \times \sum f$

$\sum fx = 1445 \times 10$

$= 14450$

10 workers salary = \$14450

11 workers salary = \$14450 + 1500 = \$15950

Average salary = 15950/11

=1450

Practice Problem -2

Example 2: The data on the number of patients who visit a hospital in a month are shown in the following table. Find out how many patients typically visit the hospital each day.

Number of patients	Number of days visiting hospital
0-10	3
10-20	7
20-30	10
30-40	8
40-50	5
50-60	3

Practice Problem -2 (Solution)

Classmark (xi)	frequency (fi)	xifi
5	3	15
15	7	105
25	10	250
35	8	280
45	5	225
55	3	165
Total	$\sum f = 36$	$\sum fx = 1040$

Mean = $x = \sum xf / \sum f = 1040/36 = 28.89$

Practice Problem -3

Example 3: A school surveyed the heights (in cm) of 50 girls in a class, and the results are provided in the manner below. Discover the data's mode.

Height	Number of girl
120-130	4
130-140	7
140-150	12
150-160	20
160-170	~~8~~ 7
Total	51

Practice Problem -3 (Solution)

Modal class = 150 - 160 [as it has maximum frequency]

$l = 150$,

$h = 10$,

$f_m = 20$,

$f_1 = 12$,

$f_2 = 7$

Mode = $l + [(f_m - f_1)/(2f_m - f_1 - f_2)] \times h$

$= 150 + [(20 - 12)/(2 \times 20 - 12 - 8)] \times 10$

$= 150 + 4$

$= 154$

Practice Problem -4

Example 4: The following are the marks scored by the students in the Summative Assessment exam. Find the median marks.

Marks	Number of girl
0-10	2
10-20	7
20-30	15
30-40	10
40-50	11
50-60	5

Practice Problem - 4 (Solution)

Marks	Number of girl	Cumulative Frequency
0-10	2	2
10-20	7	9
20-30	15	24
30-40	10	34
40-50	11	45
50-60	5	50

Median class = (N/2)th value = (50/2)th value = 25th value

Median class = 20 - 30

l = 20, N//2 = 25, m = 9, f = 15 and c = 10

Median = 20 + ([25 - 9]/15) x 10

= 20 + (16/15) x 10

= 20 + 10.6

= 30.6

≈ 31

Measure of position

In this section, we will cover :

- Deciles
- Quartiles
- Percentiles
- Boxplot
- Detecting Outlier through Box Plot
- Choosing right measures of Central Tendency

Measure of Position

We can determine where a specific data point or value falls in a sample or distribution using measures of location. A metric can help us determine whether a value is typical or out of the ordinary high or low.
Quantitative data that falls on a numerical scale is measured using position measures.
Measures can occasionally be applied to ordinal variables or variables with an order, such as first, second, and fiftieth.

Several illustrations of measure of position are:

- Cat Exam Percentile
- Top 25% of employees in the firm (Quartile)
- Top 10% of sources of revenue (decile)

Positional measurements can also demonstrate how values from various distributions or measuring scales compare.

Percentile

The value that a specific percentage falls under is known as the percentile.
For instance, if Ben is the fourth tallest child in a group of 20, then 16 out of the other 20 kids are shorter than Ben, or 80% of them. As a result, Ben is in the 80th percentile. The SAT, LSAT, and other competitive tests are where it is most frequently employed.

P = (n/N) × 100
where,
n = ordinal rank of the given observation or number of values below the observation
N = number of values in the data set
P = percentile

The ratio of values below x to all values, multiplied by 100, is known as the percentile of x.
For example, **Percentile = (Number of Values Below "x" / Total Number of Values) 100** is the percentile formula.

Percentile – Example: 1

Example 1: The scores obtained by 10 students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92. Using the percentile formula, calculate the percentile for score 70.

Solution:

Given:

Scores obtained by students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92

Number of scores below 70 = 6

Using percentile formula,

Percentile = (Number of Values Below “x” / Total Number of Values) × 100

Percentile of 70

= (6/10) × 100

= 0.6 × 100 = 60

Therefore, the percentile for score 70 = 60

Percentile – Example: 2

Example 2: The weights of 10 people were recorded in kg as 35, 41, 42, 56, 58, 62, 70, 71, 90, 77. Find the percentile for the weight 58 kg.

Solution:

Given:

Scores obtained by students are 35, 41, 42, 56, 58, 62, 70, 71, 77, 90

Number of people with weight below 58 kg = 4

Using percentile formula,

Percentile = (Number of Values Below “x” / Total Number of Values) × 100

Percentile for weight 58 kg

= (4/10) × 100

= 0.4 × 100 = 40

Therefore, the percentile for weight 58 kg = 40

Percentile – Practice Problem

Example 3: Ten students' scores are listed and announced in a college. There are 56, 45, 69, 78, 72, 94, 82, 80, 63, and 59 points. Find the 70th percentile using the percentile formula.

Percentile – Practice Problem (Solution)

Arrange the data in ascending order - 45, 56, 59, 63, 69, 72, 78, 80, 82, 94.

Find the rank.

Rank = Percentile × 100

Rank = 70 × 100

Rank = 0.07

Find the percentile using the formula:

Percentile = Rank × Total number of the data set

Percentile = 0.07 × 10

Percentile = 0.7

Since it is not a whole number, round the number to the nearest whole number i.e. 0.7 = 1

Count the set to reach the above number. It is 1.

Therefore, the 70th percentile is 45.

Decile

A decile is a quantile that uses 9 data points to divide the dataset into ten equal subsections. Each part of the sorted data represents one-tenth of the initial sample or population. Decile aids in arranging massive volumes of data in ascending or descending order. The scale used for this ordering ranges from 1 to 10, with each value after that denoting an increase of 10 percentage points.

Ungrouped Data: D(x) = (n+1) * x/10

The decile value, x, which might vary from 1 to 9, must be computed. The data set's total number of observations is n.

Decile Formula for grouped data: D(x) = l + (w/f) ((Nx)/10 - C)

w is the class's size, N is the overall frequency, and C is the cumulative frequency of the class that comes before it. L is the lower border of the class that contains the decile determined by (x cf) / 10, cf is the cumulative frequency of the complete data set, and N is the total frequency.

Decile – Example: 1

Suppose a data set consists of the following numbers: 24, 32, 27, 32, 23, 62, 45, 80, 59, 63, 36, 54, 57, 36, 72, 55, 51, 32, 56, 33, 42, 55, 30. Calculate the 6th decile and
9th Decile.

Solution: The arranged data is 23, 24, 27, 30, 32, 32, 32, 33, 36, 36, 42, 45, 51, 54, 55, 55, 56, 57, 59, 62, 63, 72, 80

n = 23

D(6) = 6(n+1)106(n+1)10 = 14.4th data. This lies between 54 and 55.

D(6) = 54 + 0.4 * (55 – 54) = 54.4

D(9) = 9(n+1)109(n+1)10 = 21.6th data. This lies between 63 and 72

D(9) = 63 + 0.6 * (72 – 63) = 68.4

Answer: D(6) = 54.4 and D(9) = 68.4

Decile – Example: 2

Find the 7th decile for the following frequency distribution table

Class	Frequency
10 - 20	15
20 - 30	10
30 - 40	12
40 - 50	8
50 - 60	7
60 - 70	18
70 - 80	5
80 - 90	25

Decile – Example: 2 (Solution)

Class	Frequency	Cumulative Frequency (cf)
10 - 20	15	15
20 - 30	10	25
30 - 40	12	37
40 - 50	8	45
50 - 60	7	52
60 - 70	18	70
70 - 80	5	75
80 - 90	25	100

$D(7) = (7\times100)/10 = 70$th data in the cf column

This data lies in the 60 - 70 class

$D(7) = 60+(10/18)*((7\times100/10)-52) = 70$

Quartile

The value that divides a list of numbers into quarters can be calculated using the quartile formula. The information is separated into quartiles after being initially organised in ascending order.
While quartiles divide data into four equal halves, the median divides it into two equal portions. Using the quartile formula, we may determine the first, second, and third quartiles.

The quartiles are shown when the observations are organised in ascending order:

1. ((n + 1)/4)th Term) is the first quartile (Q1)
2. ((n + 1)/2)th Term) is the second quartile (Q2)
3. Third Quartile (Q3) equals the (3(n + 1)/4)th term.

IQR = UPPER QUARTILE(Q3) – LOWER QUARTILE(Q1)

Quartile: Example-1

Calculate the median, lower quartile, upper quartile, and interquartile range of the following data set of values: 20, 19, 21, 22, 23, 24, 25, 27, 26

Solution:

Arranging the values in ascending order: 19, 20, 21, 22, 23, 24, 25, 26, 27

Putting the values in the formulas above, we get,

Median(Q2) = 5th Term = 23

Lower Quartile (Q1) = Mean of 2nd and 3rd term = (20 + 21)/2 = 20.5

Upper Quartile(Q3) = Mean of 7th and 8th term = (25 + 26)/2 = 25.5

IQR = Upper Quartile−Lower Quartile

IQR = 25.5 – 20.5

IQR = 5

Quartile: Example- 2

Example 2: What will be the upper quartile for the following set of numbers?
26, 19, 5, 7, 6, 9, 16, 12, 18, 2, 1.

Solution:
The formula for the upper quartile formula is Q3 = $\frac{3}{4}(n + 1)$th Term.
The formula instead of giving the value for the upper quartile gives us the place. For example, 8th place, 10th place, etc.
So firstly we put your numbers in ascending order: 1, 2, 5, 6, 7, 9, 12, 16, 18, 19, 26. There are a total of 11 numbers, so:
Q3 = $\frac{3}{4}(n + 1)$th Term.
Q3 = $\frac{3}{4}(12)$th Term. = 9th Term
Q3 = 18

What is an Outlier?

An outlier is an unusually large or small observation. Outliers can have a disproportionate effect on statistical results, such as the mean, which can result in misleading interpretations.

For example, a data set includes the values: 1, 2, 3, and 34. The mean value, 10, which is higher than the majority of the data (1, 2, 3), is greatly affected by the extreme data point, 34. In this case, the mean value makes it seem that the data values are higher than they really are. You should investigate outliers because they can provide useful information about your data or process. Often, it is easiest to identify outliers by graphing the data.

Intro to Boxplot

A box plot or box plot graphically represents groups of numerical data through their quartiles in descriptive statistics. Box-and-whisker plots and box-and-whisker diagrams refer to box plots that can additionally feature vertical lines (whiskers) demonstrating variability outside the upper and lower quartiles.

Using a five-number summary (the minimum, first quartile (Q1), median, third quartile (Q3), and "maximum"), boxplots are a common approach to depict data distribution.

Plotting a Boxplot

Access the tab raw data:

https://docs.google.com/spreadsheets/d/17Mk7wypZyEaw-nGmeRYKKTK-GOF92uFp/edit?usp=sharing&ouid=107266068801601122977&rtpof=true&sd=true

Plotting a Boxplot

Create a new tab –’Box Plot’ and calculate minimum, quartile 1, median, quartile 3, and maximum.

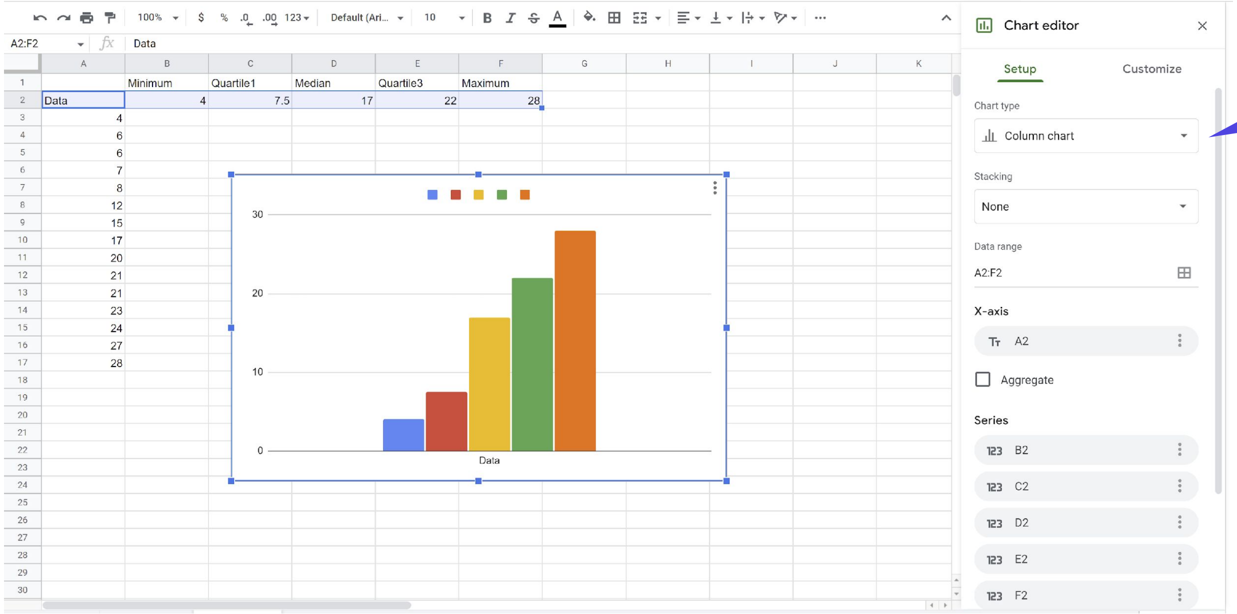
A28 fx

	A	B	C	D	E	F	G
1		Minimum	Quartile1	Median	Quartile3	Maximum	
2	Data	4	7.5	17	22	28	
3	4						
4	6						
5	6						
6	7						
7	8						
8	12						
9	15						
10	17						
11	20						
12	21						
13	21						
14	23						
15	24						
16	27						
17	28						

Raw Data | Box Plot

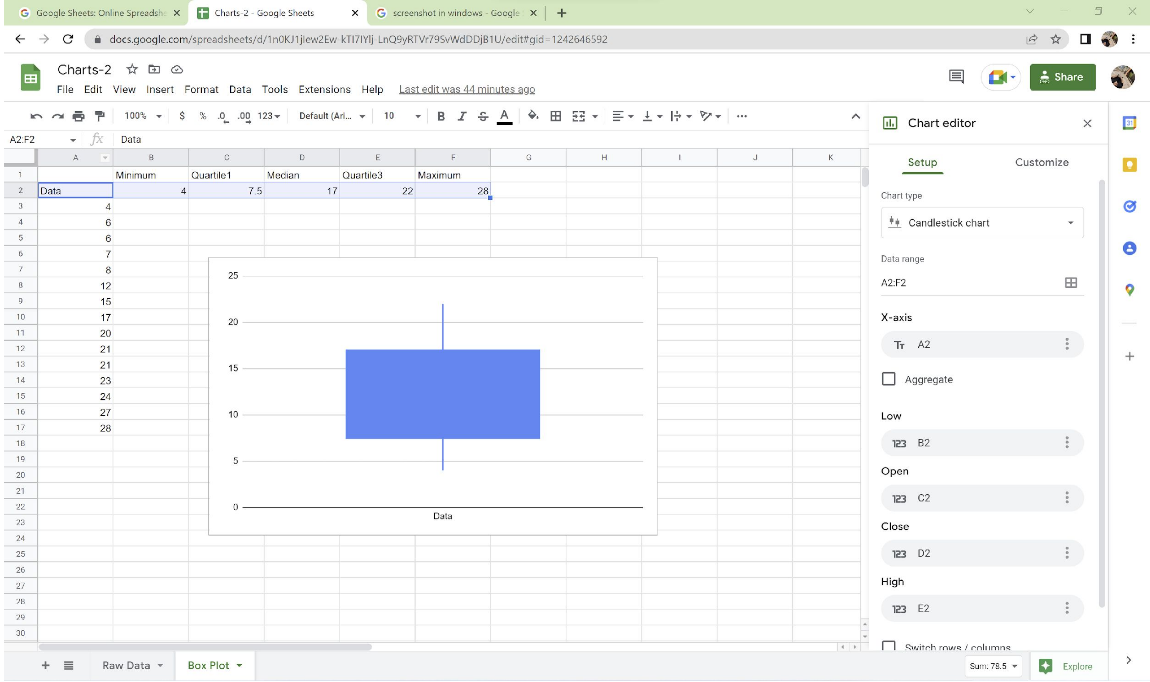
Plotting a Boxplot

Select A1:F2 and insert chart. We will get a column chart.



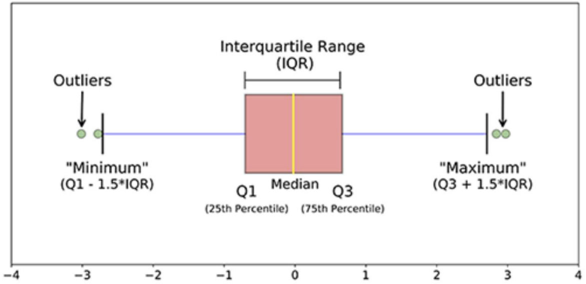
Go to chart type and change it to candlestick chart

Plotting a Boxplot



Detecting Outlier Through Boxplot

Box plot is used to detect outliers easily. It can also tell us if your data is symmetrical, how tightly your information is grouped, and if and how your data is skewed.



- 25th to 75th percentiles are the **interquartile range** (IQR). IQR reveals how widely distributed the middle values are.
- **Outliers**: "maximum": Q3 + 1.5*IQR "minimum": Q1 -1.5*IQR (shown as green circles) An observation point that is far from other observations is referred to as an outlier in statistics.

Detecting Outlier – A Practice Problem

Refer to the tab (Outlier Detection – Data) sheet:
https://docs.google.com/spreadsheets/d/17Mk7wypZyEaw-nGmeRYKKTK-GOF92uFp/edit?usp=sharing&ouid=107266068801601122977&rtpof=true&sd=true

Instruction:

1. Calculate the Q1(quartile-1), Q3(quartile-3), and IQR(Q3-Q1)
2. Create a column to identify if a data is an upper outlier(> Q3+1.5* IQR) and lower outlier(< Q1 -1.5*IQR)
3. Replace the outlier value with the upper outlier or lower outlier value

Detecting Outlier – A Practice Problem (Solution)

Refer to the tab (Outlier Detection – Data (solution) sheet:
https://docs.google.com/spreadsheets/d/1PJeO04vlZ_hDtqvFoZ6vpAQ2hkSwC-q9SZVioRH6gx4/edit?usp=sharing

Choosing Right Measures of Central Tendency

Central Tendency Measure	Pros	Cons
Mean	Sensitive as it takes all data values into account (reliable)	Biased output if outliers/ extreme values exist in the data set
Median	Not affected by extreme values	Less sensitive than Mean as it only focuses on giving out the middle data point irrespective of how far the other values are from the middle Needs the data to be arranged in the ascending order before computing
Mode	Not affected by extreme values and can be used with non-numerical data	There may be more than one mode or no mode at all and it may not reflect data summary accurately

Conclusion

THANK YOU