Statistics

Statistics is an area of applied mathematics concerned with the data collection, analysis, interpretation and presentation.

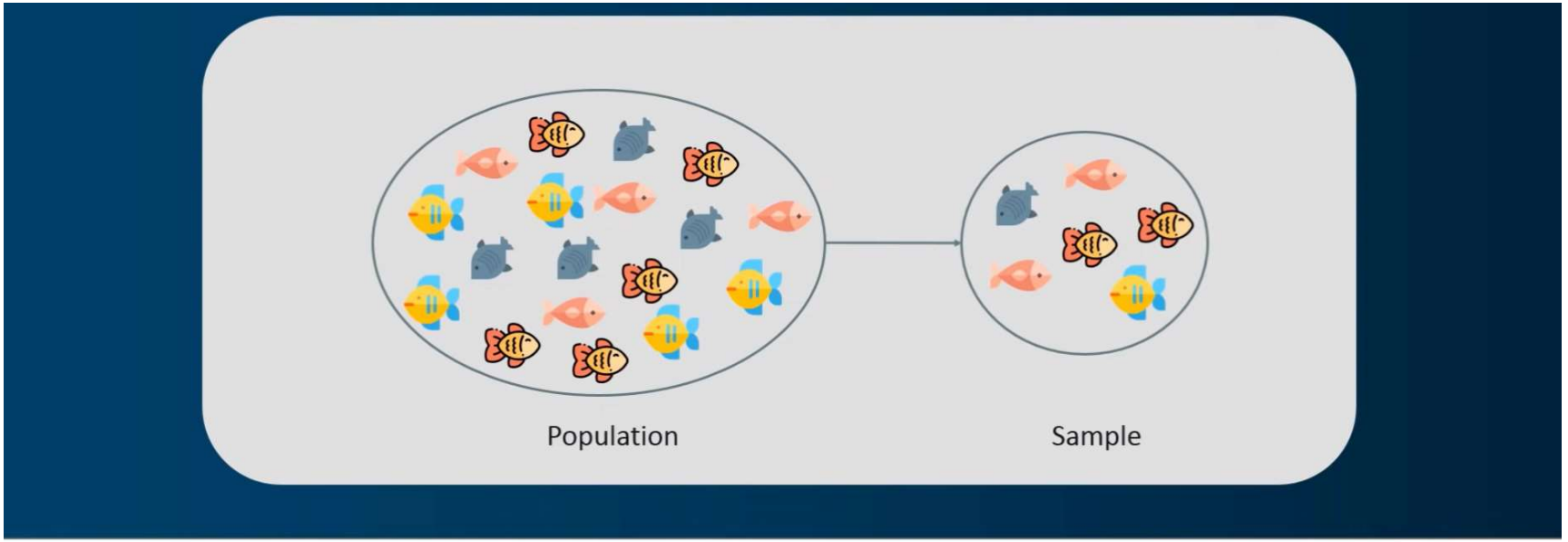
Your company has created a new drug that may cure cancer. How would you conduct a test to confirm the drug's effectiveness?

You and a friend are at a baseball game, and out of the blue he offers you a bet that neither team will hit a home run in that game. Should you take the bet?

The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you not look for?

Image Source: www.edureka.co/data-science

Statistics



Statistics Terminologies

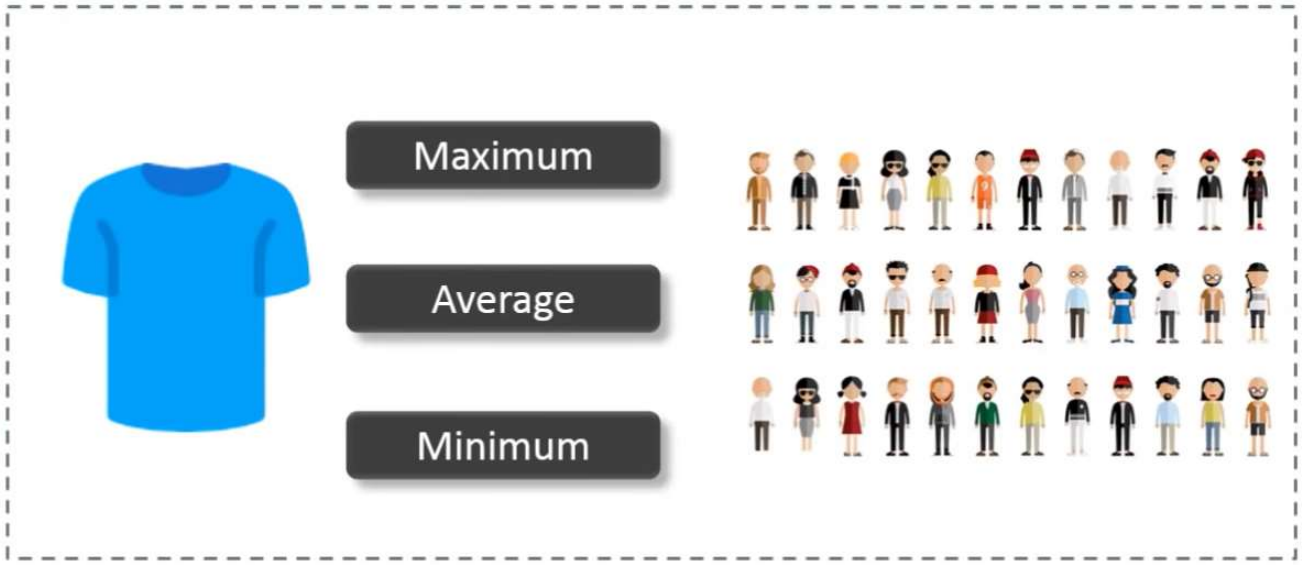
Population: A collection or set of individuals or objects or events whose properties are to be analyzed.

Sample: A subset of population is called 'Sample'. A well chosen sample will contain most of the information about a particular population parameter

Image Source: www.edureka.co/data-science

Types of Statistics: Descriptive

Descriptive statistics *uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.*



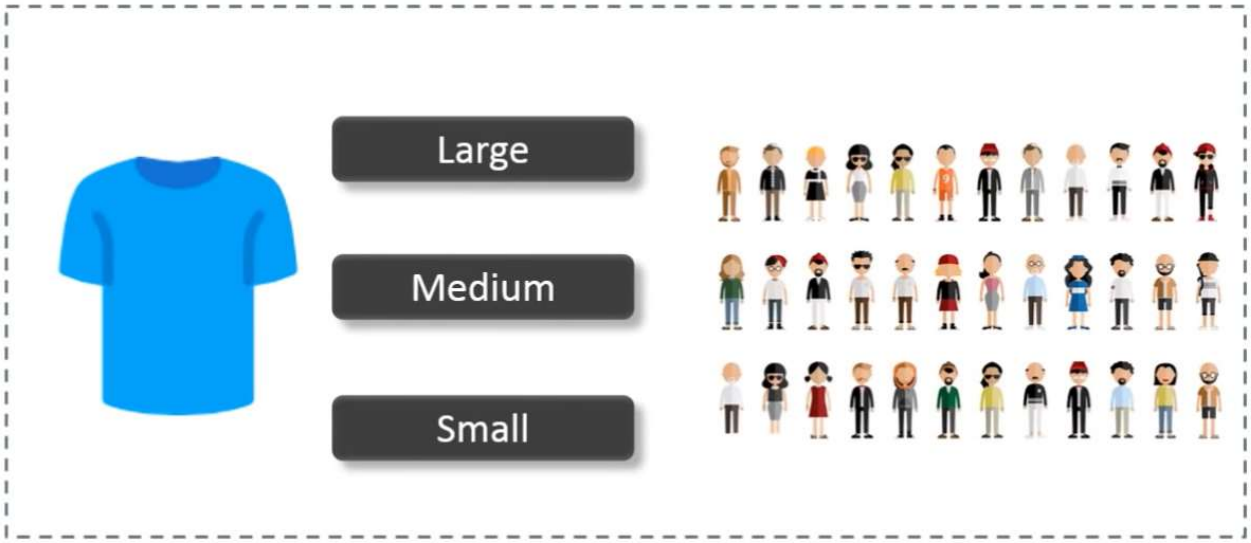
Descriptive Statistics is mainly focused upon the main characteristics of data. It provides graphical summary of the data.

Image Source: www.edureka.co/data-science

Types of Statistics: Inferential

Inferential statistics makes inferences and predictions about a population based on a sample of data taken from the population in question.

inference: a conclusion reached on the basis of evidence and reasoning.



Inferential statistics, generalizes a large dataset and applies probability to draw a conclusion. It allows us to infer data parameters based on a statistical model using a sample data.

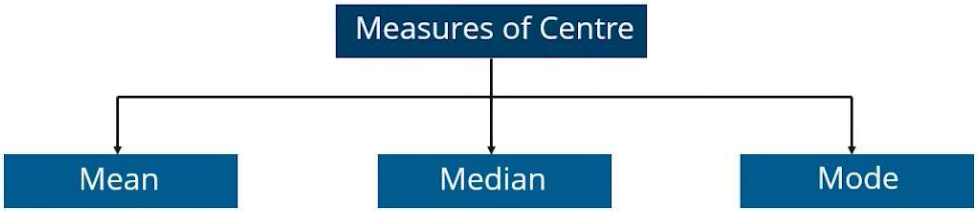
Image Source: www.edureka.co/data-science

Descriptive Statistics

Descriptive statistics are broken down into two categories:

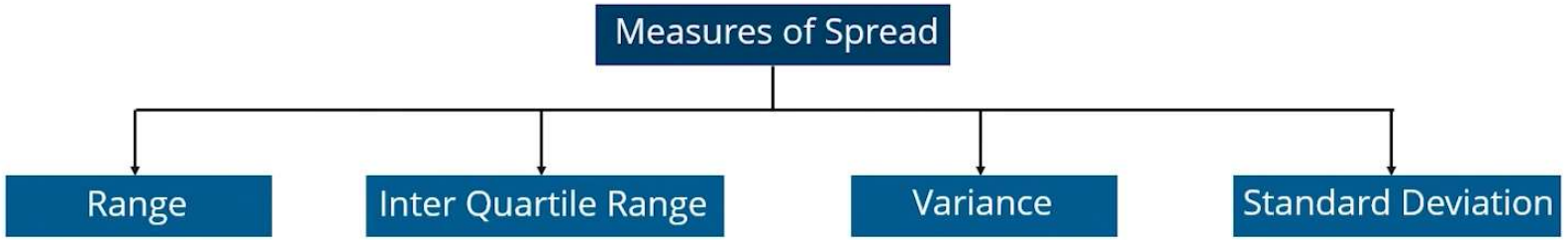
- **Measures of Central tendency**
- Measures of Variability (spread)

Image Source: www.edureka.co/data-science



Descriptive statistics are broken down into two categories:

- Measures of Central tendency
- **Measures of Variability (spread)**



Descriptive Statistics: Measures of Center Tendency

- Mean
- Median
- Mode

Mean: Measure of average of all the values in a sample

$$mean = \frac{\sum x}{n}$$

mean [10, 15, 20, 26, 28] = 19.8

Median: Measure of the central value of the sample set

Median is the central value when data is in sorted order [10, 15, **20**, 26, 28] = 20

Median [10, 15, **20**, **26**, 28, 32] = (20 + 26) / 2 = **23**

Mode: The value most recurrent in the sample set

Mode is the most frequent value [2, 5, 5, 2, 3, 2, 2, 2] = **2**

Mode is the most frequent value [2, 5, 5, 2, 3, 2, 2, 5, 2, 5, 5] = **2** or **5**

Descriptive Statistics: Measures of Center Tendency


```
import statistics
test_score = [60 , 83, 83, 91, 100]
ans = statistics.harmonic_mean(test_score)
print(f"Harmonic Mean : {ans}")
ans = statistics.mean(test_score)
print(f"Mean : {ans}")
ans = statistics.median(test_score)
print(f"Median : {ans}")
ans = statistics.median_grouped(test_score)
print(f"Grouped Median : {ans}")
ans = statistics.median_high(test_score)
print(f"High Median : {ans}")
ans = statistics.median_low(test_score)
print(f"Low Median : {ans}")
ans = statistics.mode(test_score)
print(f"Mode : {ans}")
ans = statistics.pstdev(test_score)
print(f" Population Standard deviation : {ans}")
ans = statistics.stdev(test_score)
print(f"Standard deviation : {ans}")
ans = statistics.pvariance(test_score)
print(f"Population Variance : {ans}")
ans = statistics.variance(test_score)
print(f"Variance : {ans}")
```

```
Harmonic Mean : 80.9689545753409

Mean : 83.4

Median : 83

Grouped Median : 83.25

High Median : 83

Low Median : 83

Mode : 83

Population Standard deviation : 13.275541420220872

Standard deviation : 14.842506526863986

Population Variance : 176.24

Variance : 220.3
```

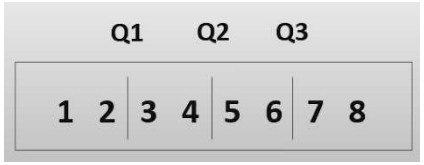
Descriptive Statistics: Measures of Spread / Dispersion

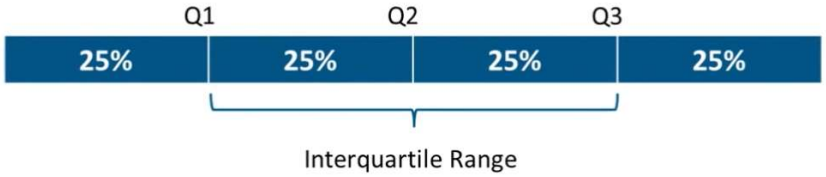
A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Range: It is measure of how spread apart the values in a datasets are. $Range = Max(X_i) - Min(X_i)$

Quartiles tell us about the spread of a dataset by breaking the dataset into quarters





Inter Quartile Range **IQR**: $Q3 - Q1$

Image Source: www.edureka.co/data-science

- Interquartile Range (IQR)

The interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile. The difference between the upper and lower quartile is known as the interquartile range.

Interquartile range = Upper Quartile – Lower Quartile
= Q3 – Q1

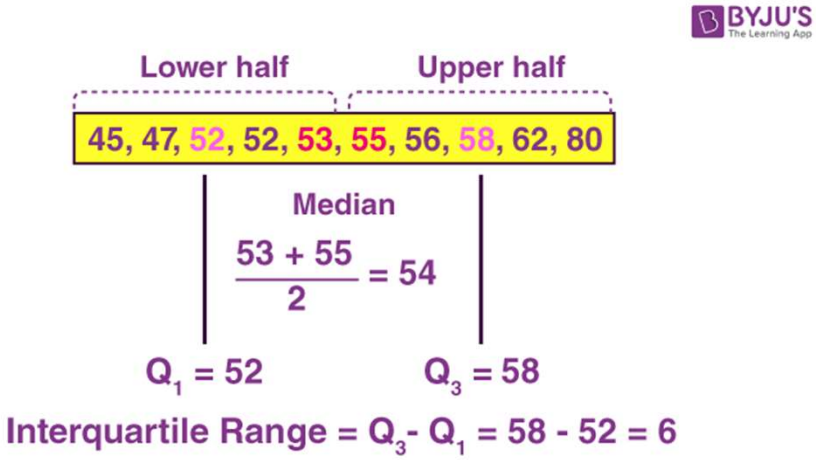


Image Source: https://byjus.com/

Descriptive Statistics: Measures of Spread / Dispersion

Consider the marks of the 100 students below, ordered from the lowest to the highest scores

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

The first quartile (Q1) lies between the 25th and 26th.
Q1 = (45 + 45) ÷ 2 = 45

The second quartile (Q2) between the 50th and 51st.
Q2 = (58 + 59) ÷ 2 = 58.5

The third quartile (Q3) between the 75th and 76th.
Q3 = (71 + 71) ÷ 2 = 71

Image Source: www.edureka.co/data-science

Descriptive Statistics: Measures of Spread / Dispersion

Variance describes how much a random variable differs from its expected value. It entails computing squares of deviations.

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

x : Individual data points
n : Total number of data points
$\bar{x}$: Mean of data points

Population Variance is the average of squared deviations.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} = (x_i - \mu)^2$$

Deviation is the difference between each element from the mean.

Deviation = $(x_i - \mu)$

Sample Variance is the average of squared differences from the mean.

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

Standard Deviation is the measure of the dispersion of a set of data from its mean.

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

Descriptive Statistics: Measures of Spread / Dispersion

It refers to how "spread out" a group of values is.

- Range
- Variance and standard deviation
- Coefficient of variation
- Interquartile Range (IQR)

Range: Largest value – Smallest Value [10, 25, 12, 13, 15, 16, 23]. Range 25−10 = **15**

Variance: How far a set of numbers is spread out from their average values [δ^2]

Standard Deviation: Standard or typical difference between each data point and the mean [δ]

Population Standard Deviation (SD): $\sqrt{\frac{\sum(X-\bar{X})^2}{N}}$

Descriptive Statistics: Measures of Spread / Dispersion

Data: [2, 7, 3, 12, 9]

Mean = [2+7+3+12+9] / 5 = 6.6

$(2-6.6)^2 + (7-6.6)^2 + (3-6.6)^2 + (12-6.6)^2 + (9-6.6)^2$ = 21.16 + 0.16 + 12.96 + 29.16 + 5.76 = 69.2

Variance = δ^2 = 69.2 / 5 = 13.84

Standard Deviation = $\delta = \sqrt{13.84}$ = 3.72

	A	B	C
1		**Data Set 1**	**Data Set 2**
2		10	100
3		12	19
4		15	150
5		8	800
6		20	659
7		18	11
8		25	239
9		13	789
10		19	1000
11	**Variance (Popuated)**	26.02	134467.36
12	**Variance (Standard)**	29.28	151275.78
13	**Standard Deviation (P)**	5.10	366.70
14	**Standard Deviation (S)**	5.41	388.94
15			

```
=VAR.P(B2:B10)
=VAR.S(B2:B10)
=STDEV.P(B2:B10)
=STDEV.S(B2:B10)
```

Descriptive Statistics: Measures of Spread / Dispersion

Population Standard Deviation (for all values)= $\sqrt{\frac{\sum (x-\mu)^2}{N}}$, μ = Population Mean

Sample Standard Deviation (for sample/selected values) = $\sqrt{\frac{\sum (x-\bar{x})^2}{N-1}}$ $\bar{x}$ = Population Mean

	A	B	C
1		**Data Set 1**	**Data Set 2**
2		10	100
3		12	19
4		15	150
5		8	800
6		20	659
7		18	11
8		25	239
9		13	789
10		19	1000
11	**Variance (Popuated)**	26.02	134467.36
12	**Variance (Standard)**	29.28	151275.78
13	**Standard Deviation (P)**	5.10	366.70
14	**Standard Deviation (S)**	5.41	388.94

```
=VAR.P(B2:B10)
=VAR.S(B2:B10)
=STDEV.P(B2:B10)
=STDEV.S(B2:B10)
```

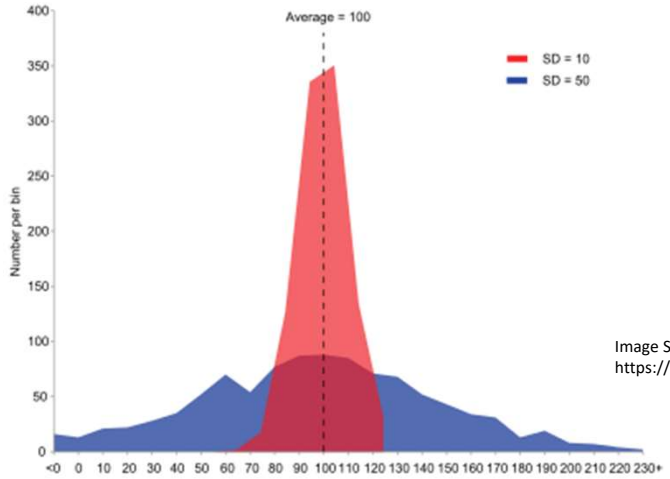


Image Source:
https://en.wikipedia.org/wiki/Variance

Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

Descriptive Statistics: Measures of Spread / Dispersion

- To compare two datasets (Relative Standard Deviation)

Coefficient of variation = $\frac{Standard\ Deviation}{Mean}$ x 100

	X	(X-μ)	$(X-\mu)^2$
	7	2	4
	2	-3	9
	7	2	4
	6	1	1
	5	0	0
	6	1	1
	2	-3	9
Total	35		28
Average μ	5		

Population Standard Deviation = $\sqrt{\frac{\sum (x-\mu)^2}{N}} = \sqrt{\frac{28}{7}} = 2$

Sample Standard Deviation = $\sqrt{\frac{\sum (x-\mu)^2}{N-1}} = \sqrt{\frac{28}{6}} = 2.16$

Coefficient of variation = $\frac{Standard\ Deviation}{Mean}$ x 100

= $\frac{2}{5}$ x 100 = 40%

Sample Coefficient of variation = $\frac{Populatio\ Standard\ Deviation}{Mean}$ x 100 = $\frac{2.16}{5}$ x 100 = 43.2%