Statistics and Probability

Hafara Firdausi, M.Kom.

Department of Information Technology
Faculty of Electrical and Intelligent Informatics
Institut Teknologi Sepuluh Nopember

02. Concepts in Statistics

Konsep Statistika

Outline

1. What is Statistics?

2. Sample and Population

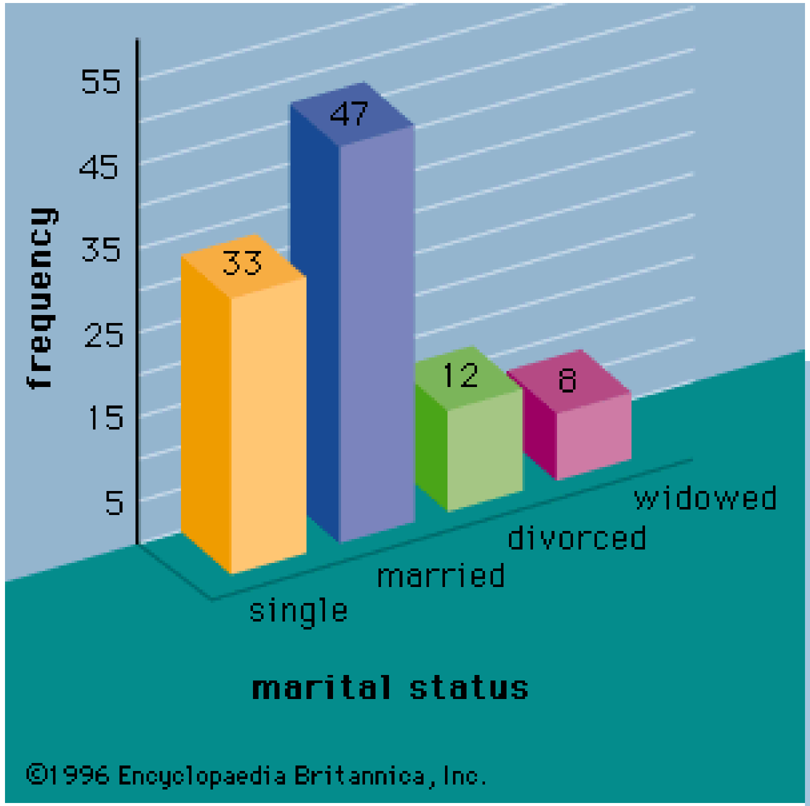
3. Summary Measures

What is Statistics?

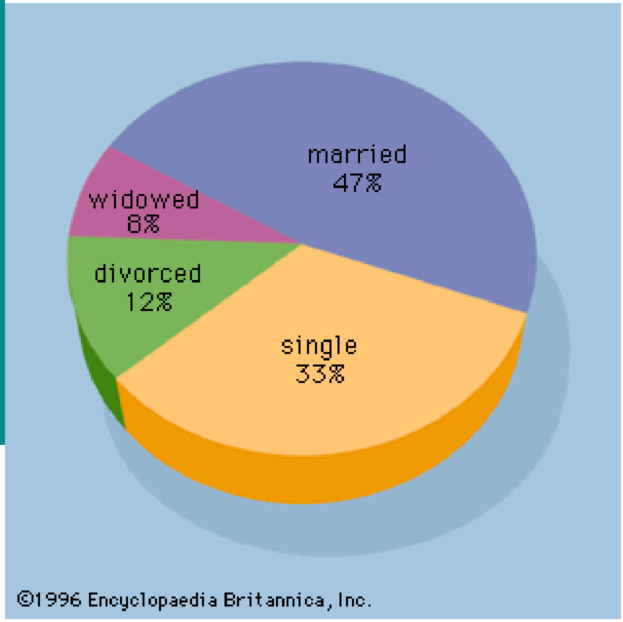
What is Statistics?

- Statistics is a **branch of mathematics** that consists of a set of **analytical techniques** that can be **applied to data** to help us make judgments and decisions in problems involving **uncertainty**.

- Statistics is a scientific discipline consisting of procedures for
  - **collecting**,
  - **describing**,
  - **analyzing**, and
  - **interpreting** numerical data



©1996 Encyclopaedia Britannica, Inc.

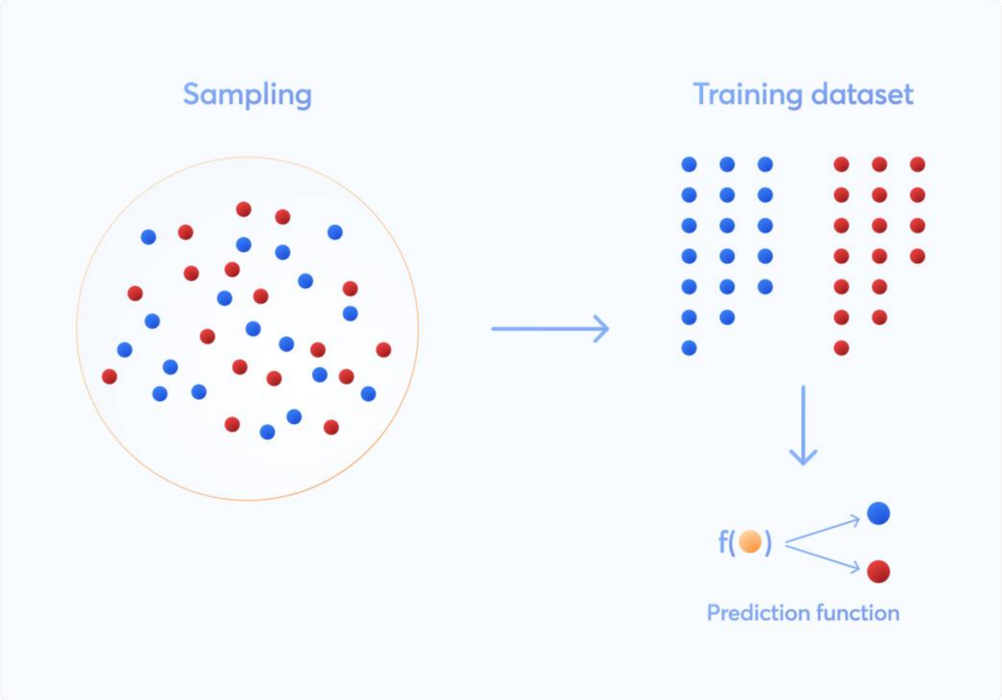


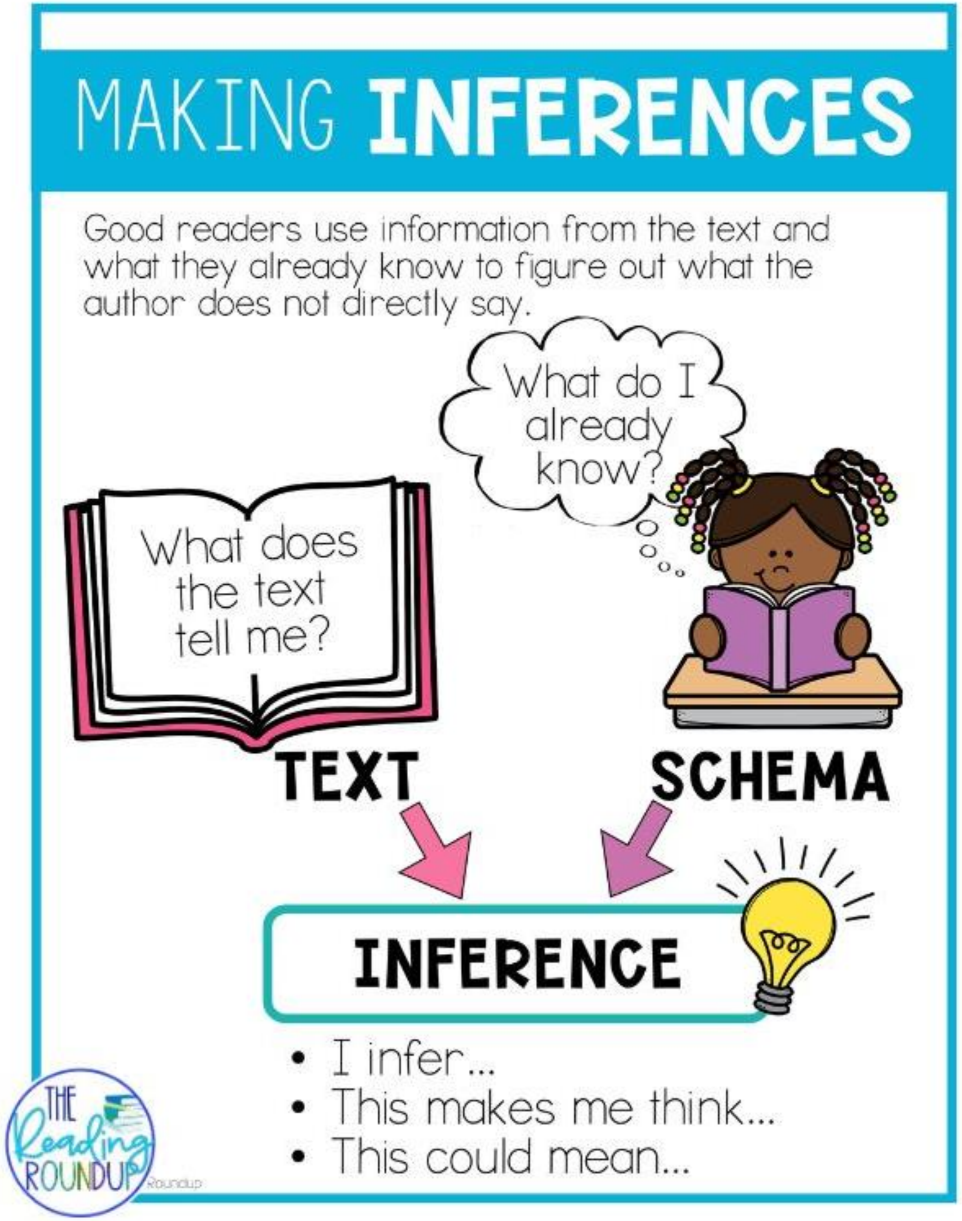
©1996 Encyclopaedia Britannica, Inc.

What is Statistics?

Main Objectives:

To provide a set of procedures that enables us to **make inferences, predictions, and decisions** about **characteristics of a population of data** based on the information obtained from only a part of the population (**sample**).

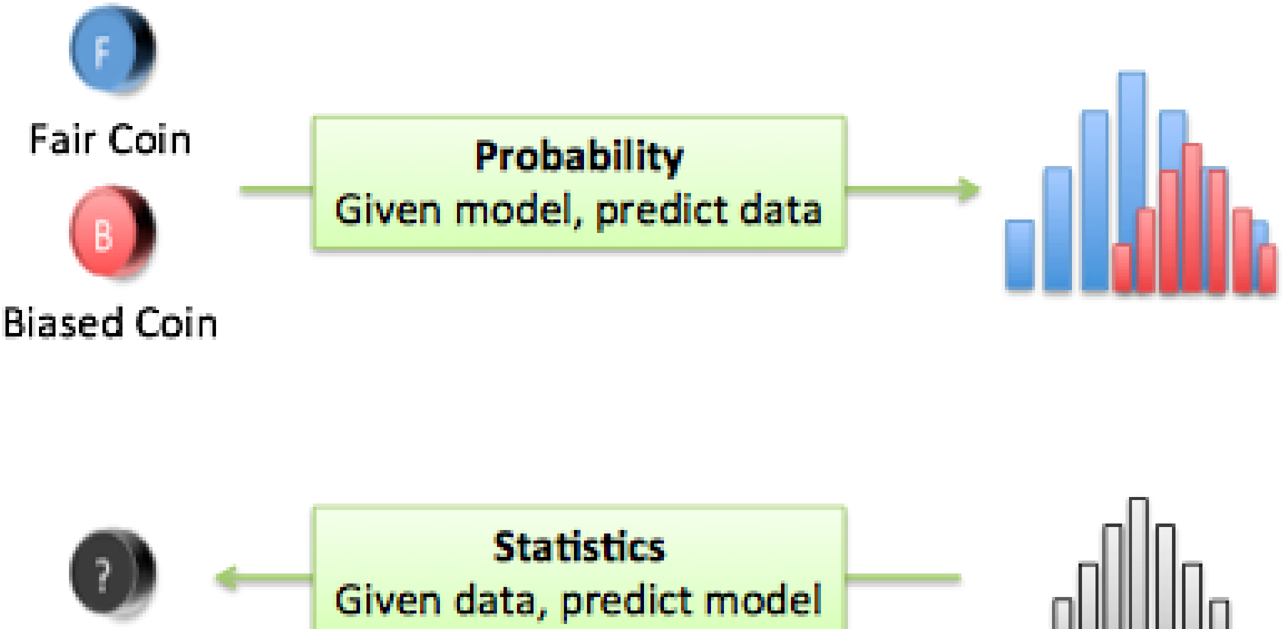



MAKING INFERENCES
Good readers use information from the text and what they already know to figure out what the author does not directly say.
What do I already know?
What does the text tell me?
TEXT
SCHEMA
INFERENCE
• I infer...
• This makes me think...
• This could mean...
THE Reading ROUNDUP

What is the Relation between Statistics and Probability?

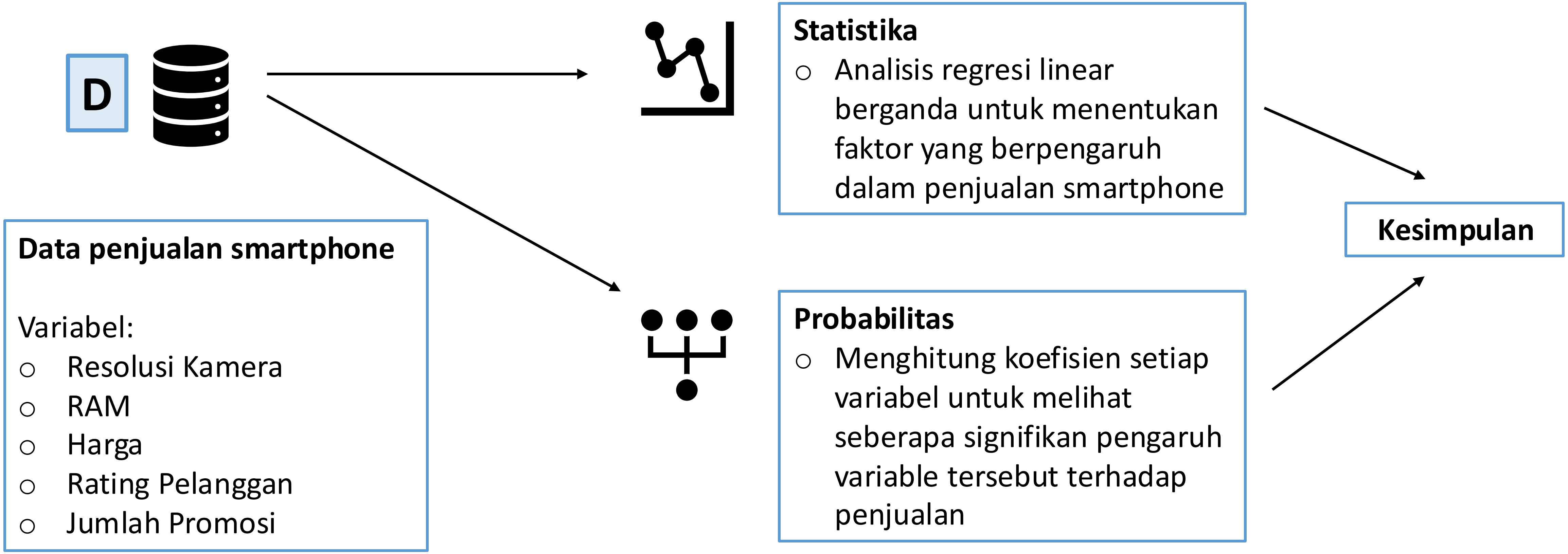
"**Probability** provides a mathematical framework for **measuring uncertainty**"

"**Statistics** uses data to draw **conclusions** that are supported by that probability"



Case Study

Analisis faktor paling berpengaruh dalam penjualan Smartphone



Data penjualan smartphone

Variabel:

- Resolusi Kamera
- RAM
- Harga
- Rating Pelanggan
- Jumlah Promosi

Statistika

- Analisis regresi linear berganda untuk menentukan faktor yang berpengaruh dalam penjualan smartphone

Probabilitas

- Menghitung koefisien setiap variabel untuk melihat seberapa signifikan pengaruh variable tersebut terhadap penjualan

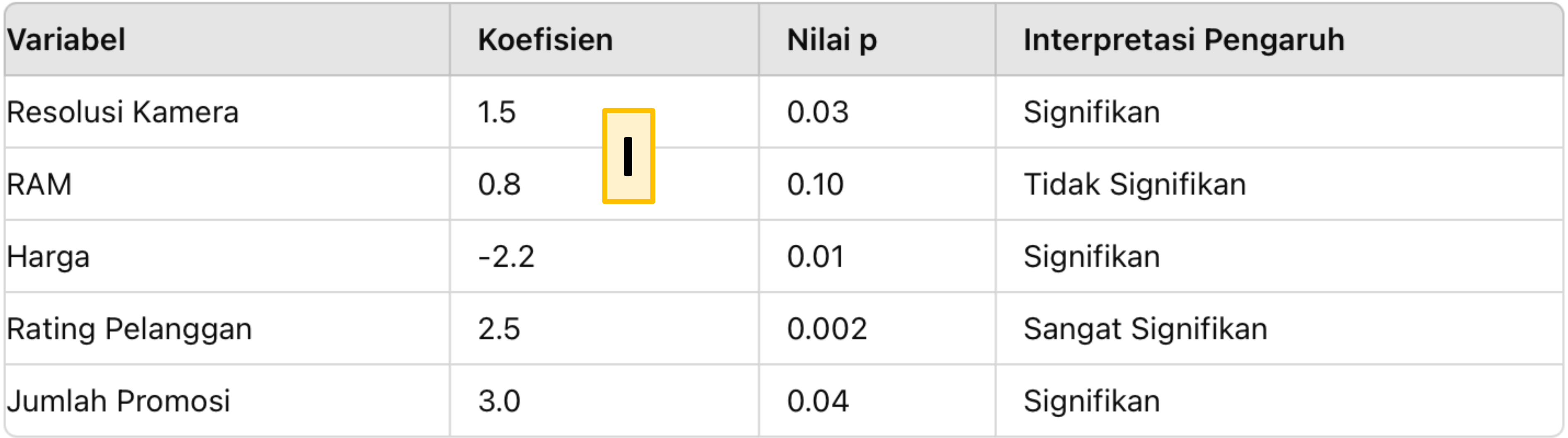
Kesimpulan

Case Study

Analisis faktor paling berpengaruh dalam penjualan Smartphone

Statistika

Probabilitas



Variabel	Koefisien	Nilai p	Interpretasi Pengaruh
Resolusi Kamera	1.5	0.03	Signifikan
RAM	0.8	0.10	Tidak Signifikan
Harga	-2.2	0.01	Signifikan
Rating Pelanggan	2.5	0.002	Sangat Signifikan
Jumlah Promosi	3.0	0.04	Signifikan

Case Study

Analisis faktor paling berpengaruh dalam penjualan Smartphone

K

Kesimpulan:

- Rating pelanggan paling berpengaruh terhadap penjualan smartphone, artinya pelanggan lebih cenderung membeli produk yang memiliki rating tinggi dan ulasan positif.
- Harga yang lebih rendah dan jumlah promosi yang banyak dapat meningkatkan penjualan.
- Resolusi kamera memiliki pengaruh lebih besar dibandingkan RAM.

Case Study

Analisis faktor paling berpengaruh dalam penjualan Smartphone

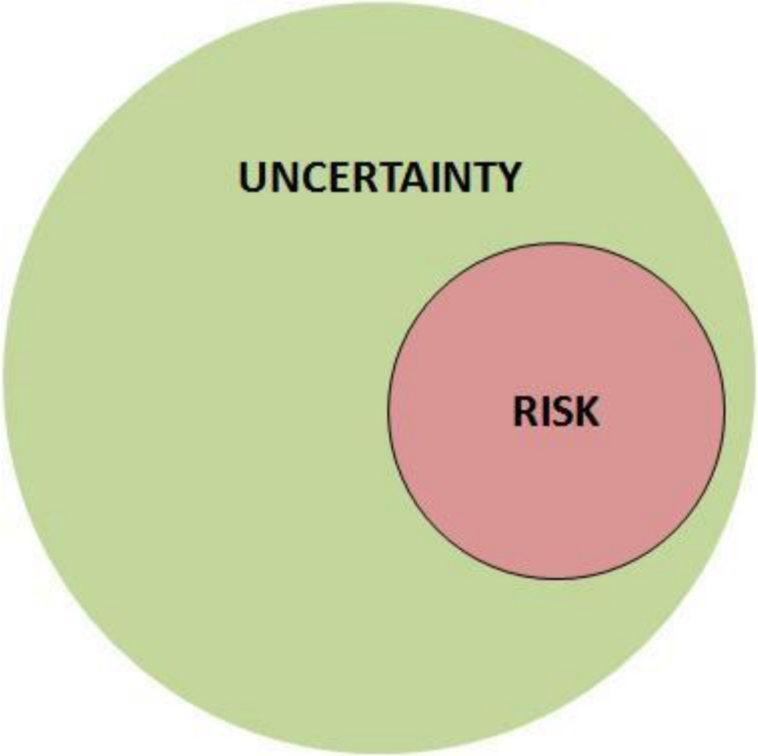
W

Keputusan:

- Meningkatkan ulasan pelanggan dan promosi untuk meningkatkan penjualan, misalnya dengan menggunakan jasa para *influencer* sosial media agar meng-*endorse* produk kita.
- Menetapkan harga yang kompetitif untuk mendorong lebih banyak penjualan.
- Memfokuskan promosi pada fitur yang lebih dihargai oleh pelanggan seperti resolusi kamera, dibandingkan RAM.

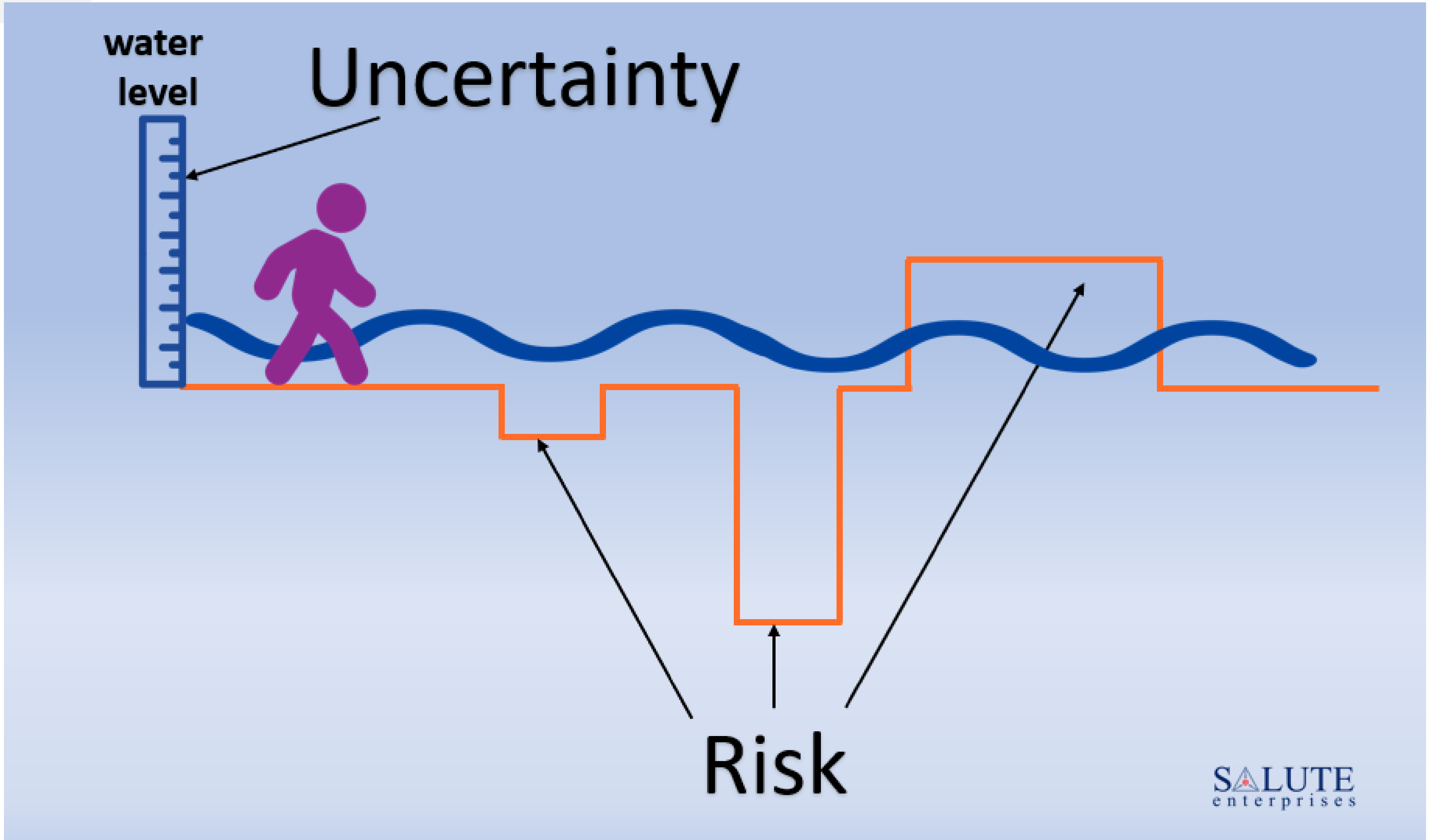
Why learn Statistics?

- **Numerical information is everywhere** and we dealing with **uncertainty**
- Statistical techniques are used to **make decisions** that affect our daily lives
- The knowledge of statistical methods will help you understand how decisions are made and give you a **better understanding** of how they affect you



This Photo by Unknown Author is licensed under CC BY-NC-ND

Why learn Statistics?



Statistics in Information Technology

- **Data Mining** is the analysis of information in a database, using tools that look for trends or irregularities in large data sets.
- **Data Compression** is the coding of data using compact formulas, called algorithms, and utilities to save storage space or transmission time.
- **Speech Recognition** is the identification of spoken words by a machine. The spoken words are turned into a sequence of numbers and matched against coded dictionaries.
- **Vision and Image Analyses** use statistics to solve contemporary and practical problems in computer vision, image processing, and artificial intelligence.
- **Human/Computer Interaction** uses statistics to design, implement, and evaluate new technologies that are useable, useful, and appealing to a broad cross-section of people.
- **Network/Traffic Modeling** uses statistics to avoid network congestion while fully exploiting the available bandwidth.
- **Stochastic Optimization** uses chance and probability models to develop the most efficient code for finding the solution to a problem.

The field of statistics

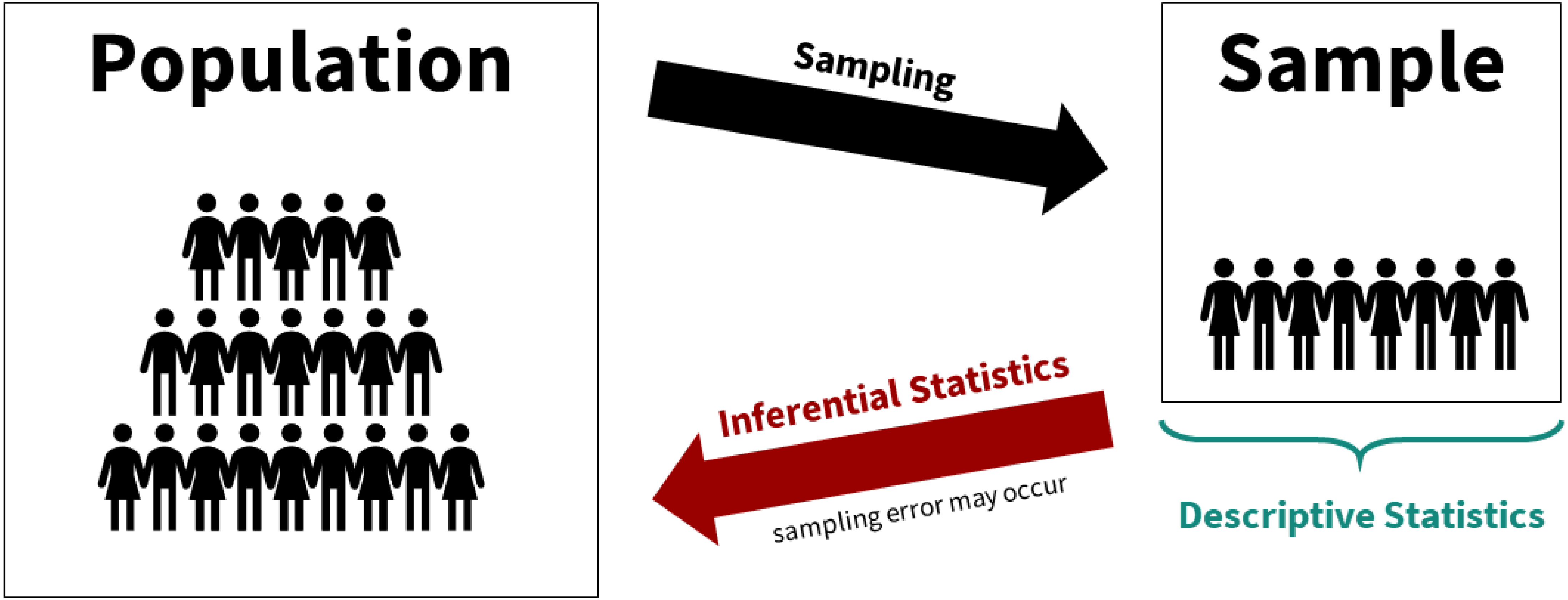
Descriptive statistics

The methods used to **summarize** quantitative and qualitative features in a sample or population and **present** data in an **informative way**.

Inferential statistics

The methods used for the **derivation of conclusions** about a population from information in a random sample of that population.

The Field of Statistics

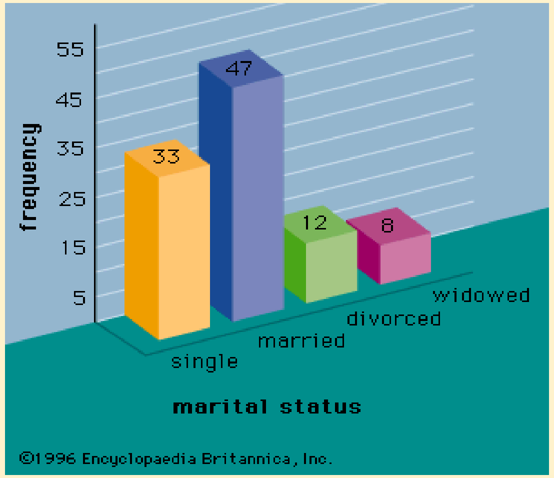


Descriptive Statistics

Descriptive statistics consist of procedures for:

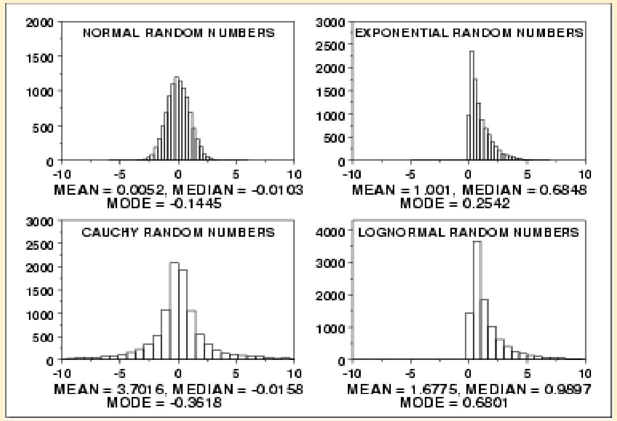
1

Tabulating or graphing the general characteristics of a set of data.



2

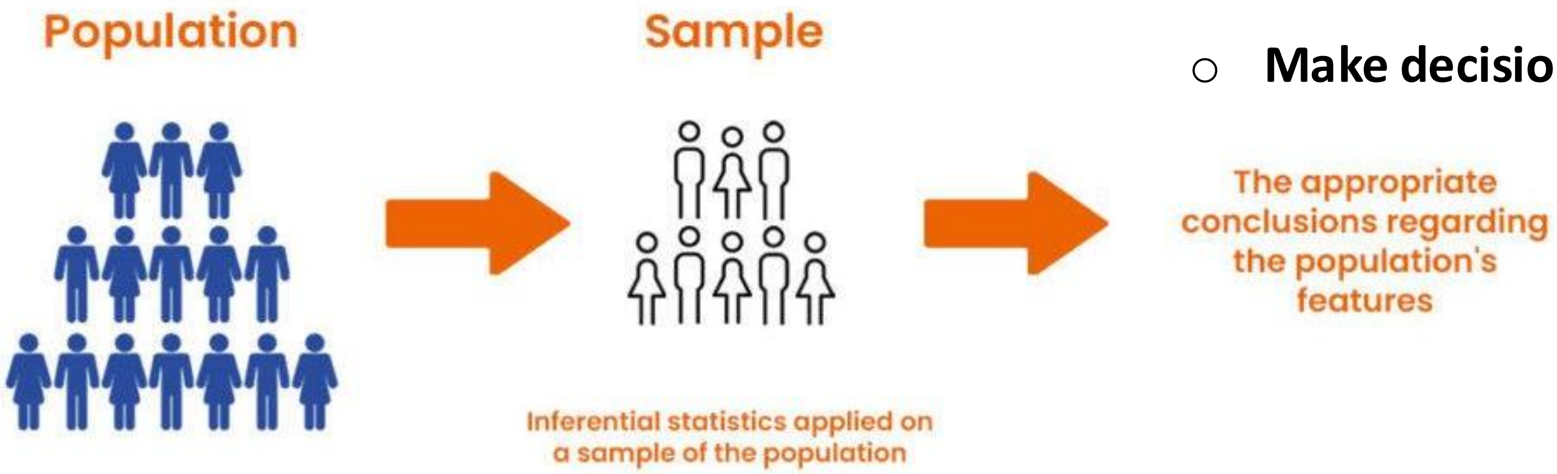
Describing some characteristics of this set such as measures of central tendency or measures of dispersion.



Inferential Statistics

Inferential statistics consists of a set of procedures that helps us make **inferences** and **predictions** about a whole population based on information from a **sample of the population**.

- ○ **Estimation / Prediction / Forecasting**
  Ex: Estimate the population mean weight using the sample mean weight
- ○ **Hypothesis testing**
  Ex: Test the claim that the population means the weight is 120 pounds
- ○ **Make decisions**



Population

Sample

Inferential statistics applied on a sample of the population

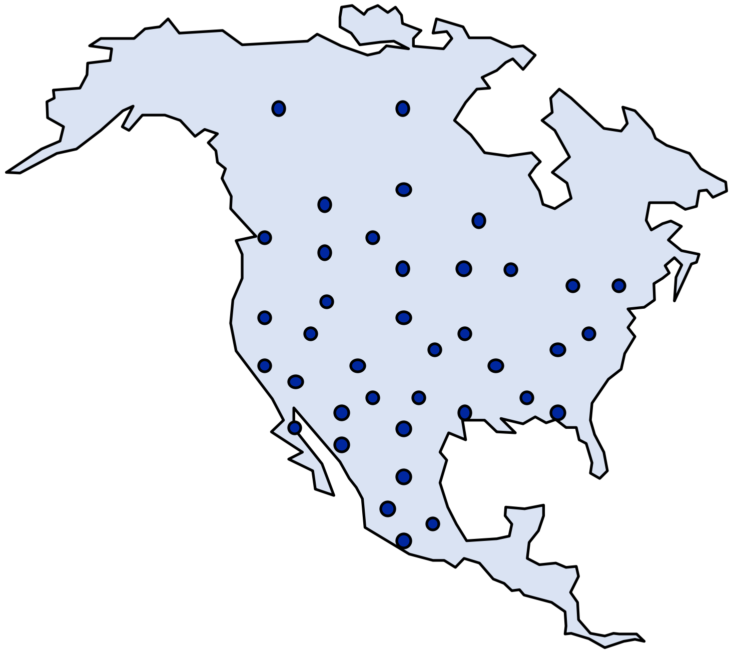
The appropriate conclusions regarding the population's features

Sample and Population

Sample and Population

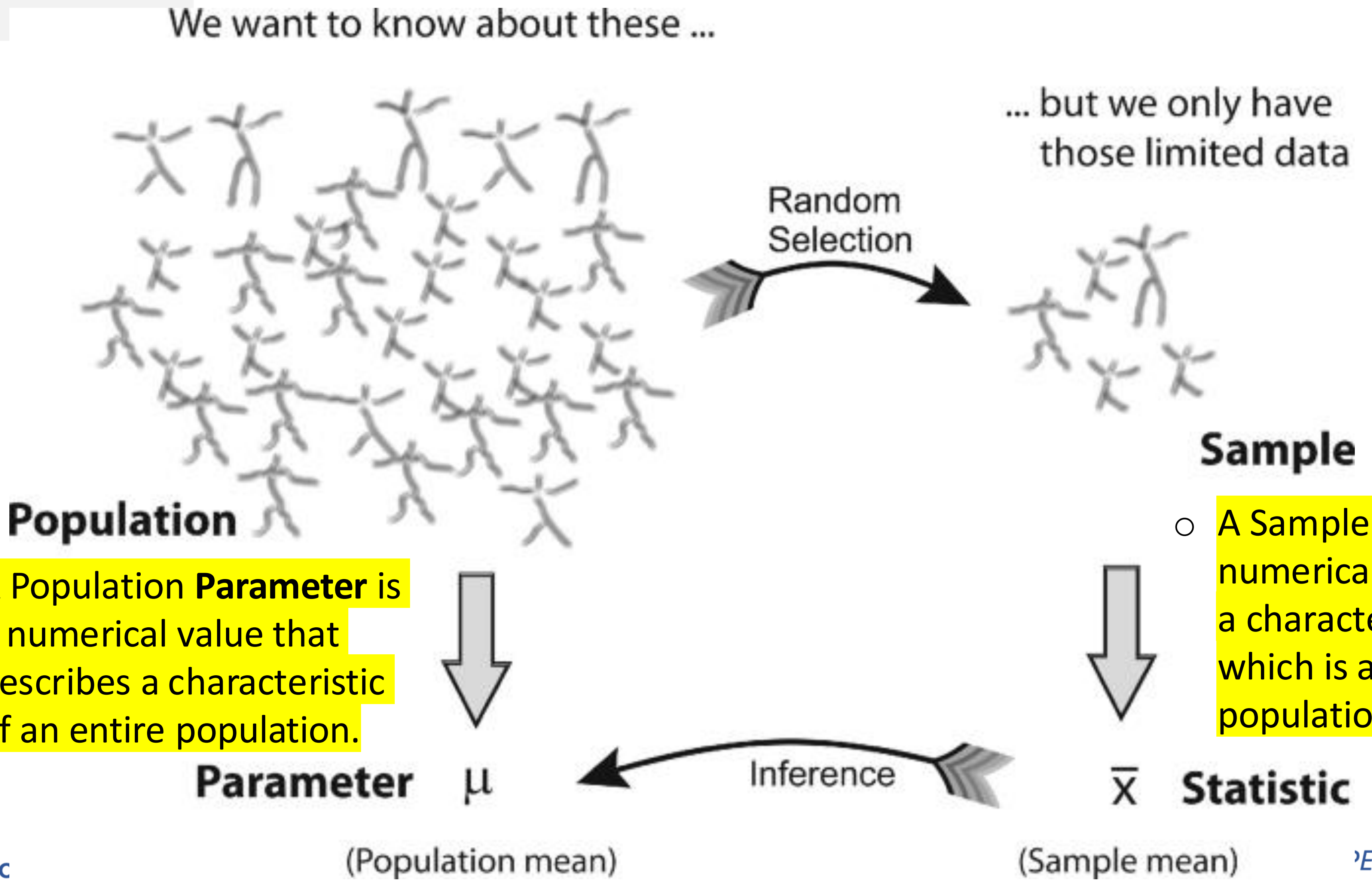
- A **population** consists of the set of **all measurements** for which the investigator is interested.
- A **sample** is a **subset of the measurements** selected from the population.
- A **census** is a **complete enumeration** of every item in a population.

Population (N)



Sample (*n*)

Population Parameter and Sample Statistic

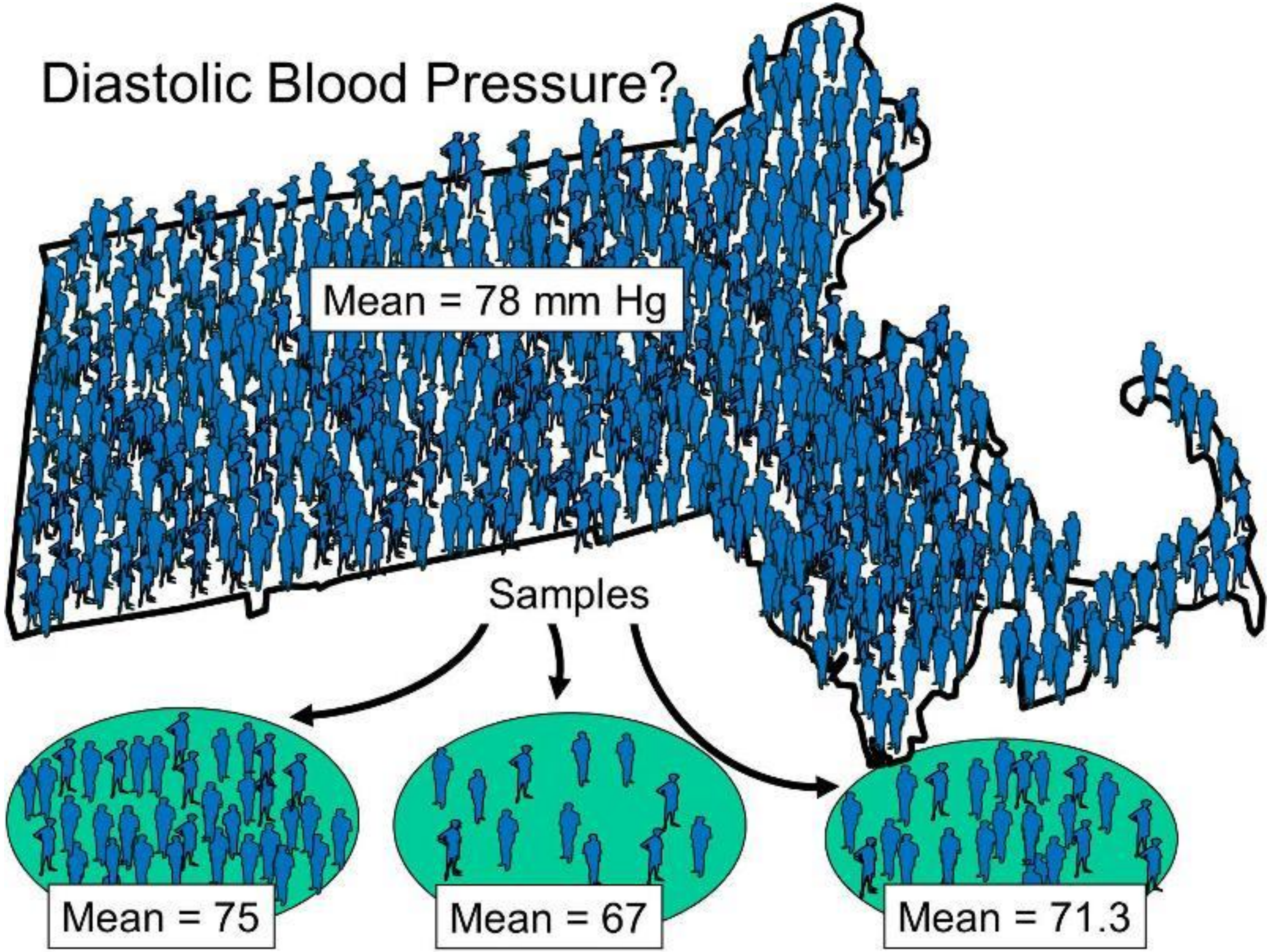


- A Population **Parameter** is a numerical value that describes a characteristic of an entire population.

- A Sample **Statistic** is a numerical value that describes a characteristic of a sample, which is a subset of the population.

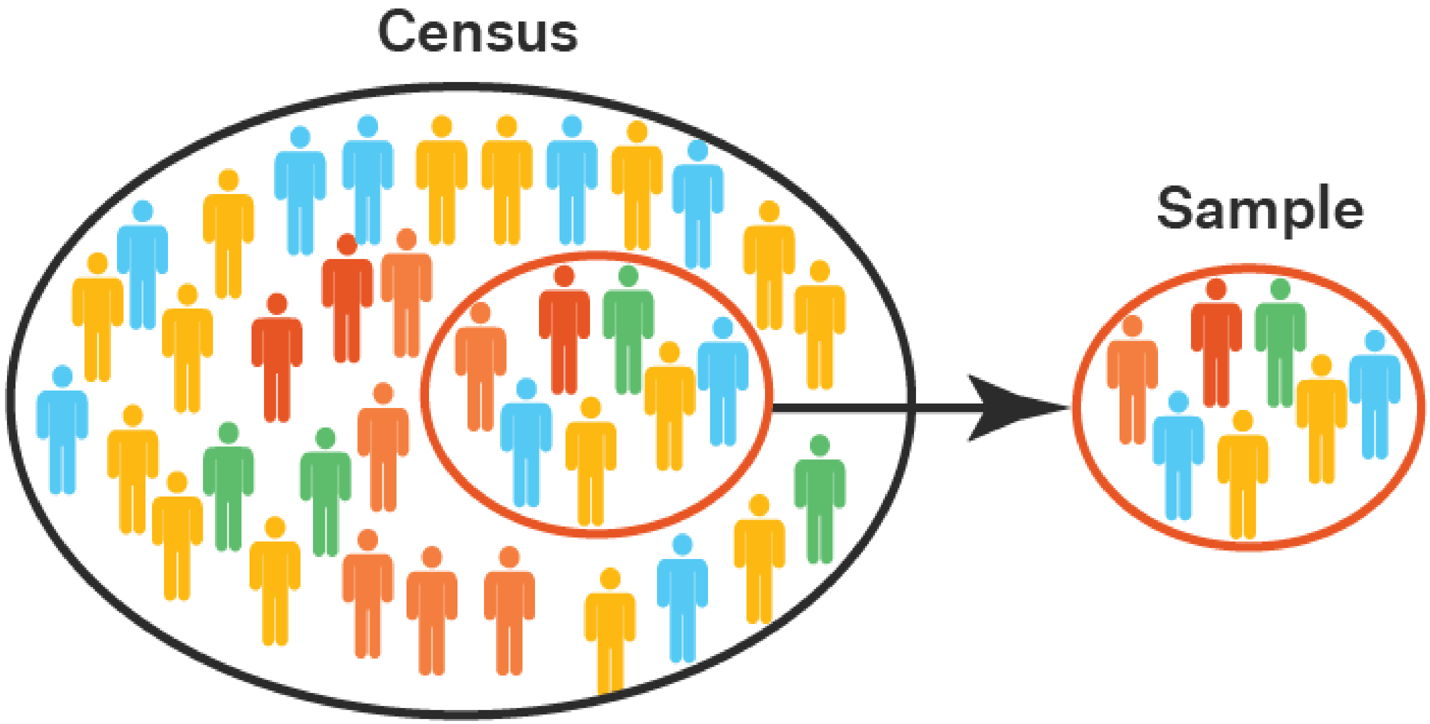
Random Sample

- **Sampling** from the population is often done **randomly**, such that every possible sample of equal size (n) will have an equal chance of being selected.
- A sample selected in this way is called a simple random sample or just a **random sample.**
- A random sample allows the chance to determine its elements.

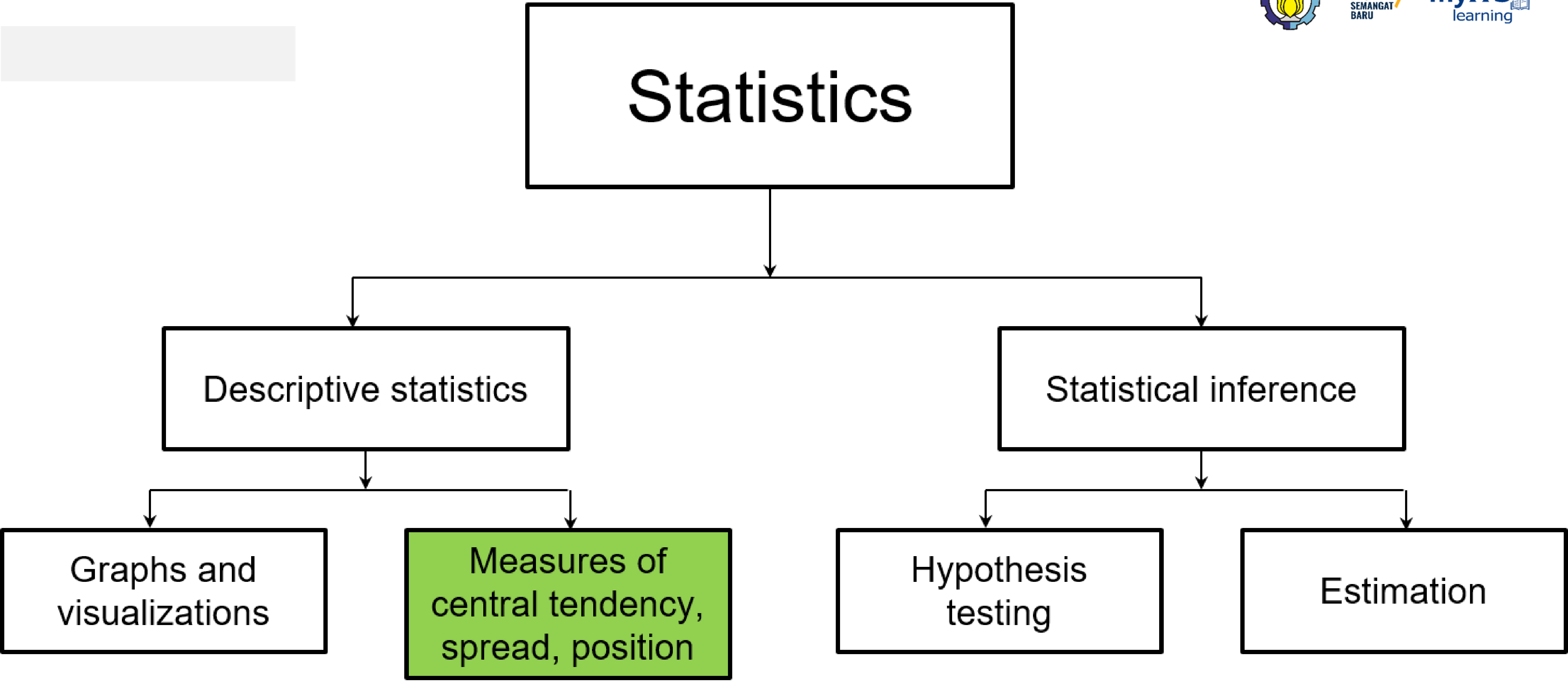


Why Sample?

- **Census** of a population may be:
    - Impossible
    - Impractical
    - Too costly



Summary Measures


Statistics

Descriptive statistics

Statistical inference

Graphs and visualizations

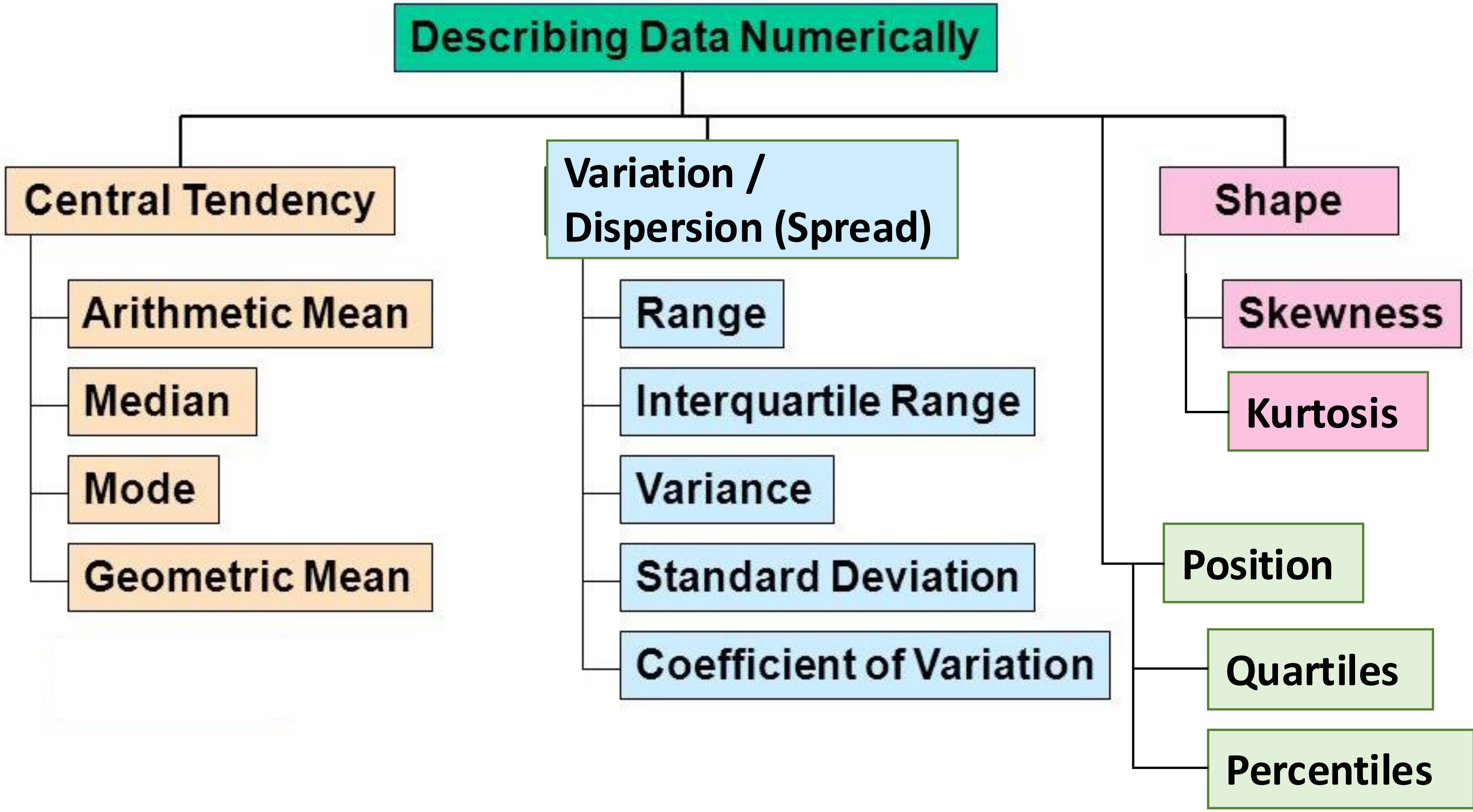
Measures of central tendency, spread, position

Hypothesis testing

Estimation

Summary Measures

- Statistical values that **summarize** or **describe key characteristics** of a dataset.
- Provide a quick, overall understanding of the data by focusing on aspects such as central tendency, spread, and distribution.



Measures of Central Tendency or Location

- Statistical tools are used to determine the **center** or **typical value** in a dataset.

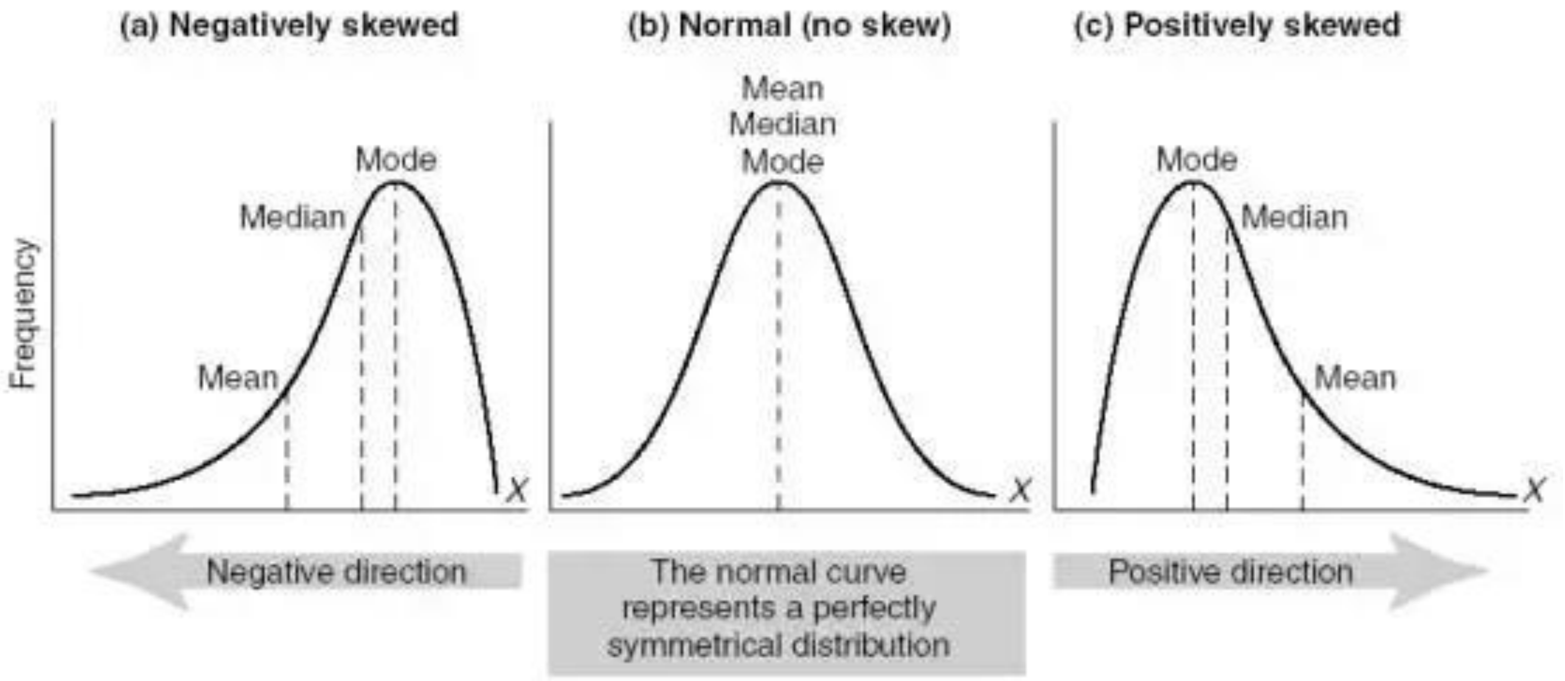
- Summarize the data by identifying a **central point** that best represents the **distribution** of values.

Median	○ Middle value when sorted in order of magnitude ○ 50th percentile
Mode	○ Most frequently-occurring value
Mean	○ Average

Measures of Central Tendency or Location

- Statistical tools are used to determine the **center** or **typical value** in a dataset.

- Summarize the data by identifying a **central point** that best represents the **distribution** of values.



Example – Median

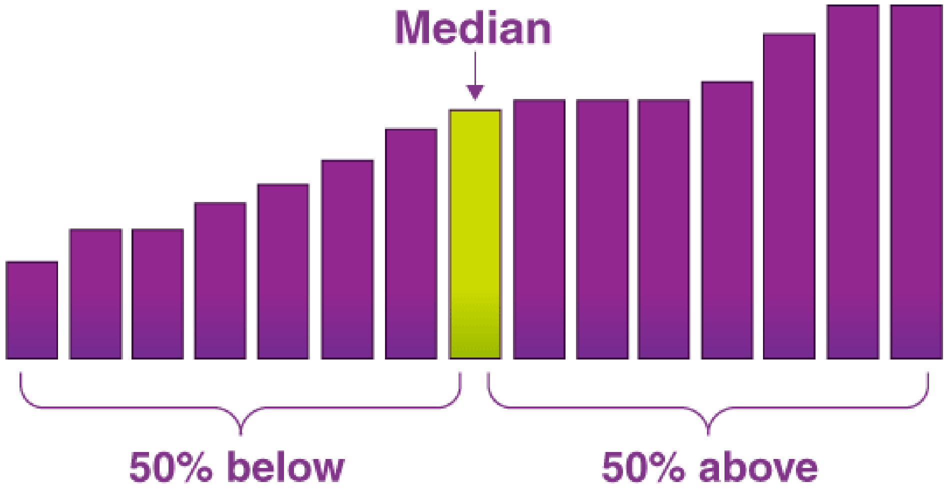
Sales	Sorted Sales	
9	6	
6	9	
12	10	
10	12	
13	13	
15	14	
16	14	
14	15	
14	16	
16	16	← Median
17	16	
16	17	
24	17	
21	18	
22	18	
18	19	
19	20	
18	21	
20	22	
17	24	

Median
50th Percentile

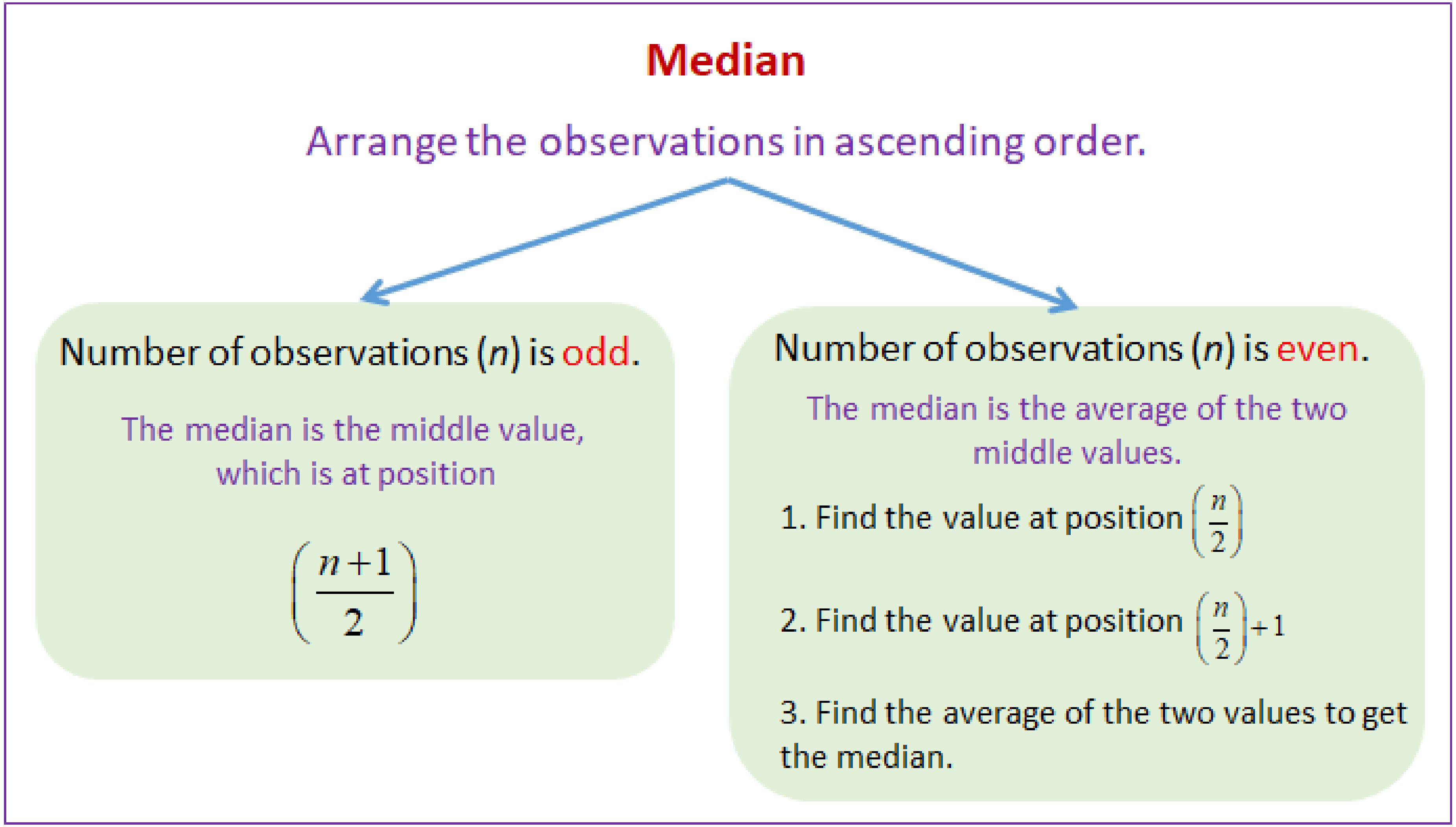
(20+1)50/100=10.5

16 + (.5)(0) = 16

- ○ The **median** is the middle value of data sorted in order of magnitude. It is the 50th percentile.
- ○ Useful for **skewed distributions or datasets with outliers**, as it isn't affected by extreme values.



Example – Median



Median

Arrange the observations in ascending order.

Number of observations (n) is odd.

The median is the middle value, which is at position

$$\left(\frac{n+1}{2}\right)$$

Number of observations (n) is even.

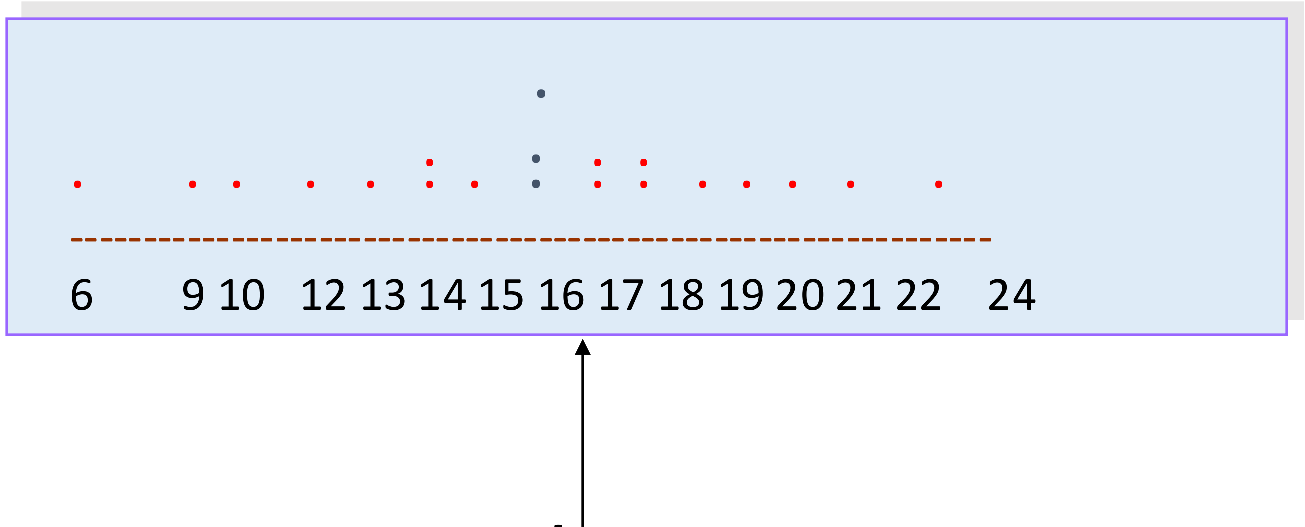
The median is the average of the two middle values.

1. Find the value at position $\left(\frac{n}{2}\right)$
2. Find the value at position $\left(\frac{n}{2}\right)+1$
3. Find the average of the two values to get the median.

Example – Mode

- The **mode** is the most frequently occurring value. It is the value with the **highest frequency**.

- Best for **categorical data** or when identifying the most common occurrence is important.



Mode = 16

Arithmetic Mean or Average

- The **mean** of a set of observations is their average - the sum of the observed values divided by the number of observations.
- Best for datasets with **evenly distributed data without extreme outliers**

Population Mean	Sample Mean
$\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	*n* = number of items in the sample

Example – Mean

Sales
9
6
12
10
13
15
16
14
14
16
17
16
24
21
22
18
19
18
20
17
317

$$x = \frac{\sum_{i=1} x}{n} = \frac{317}{20} = 15.85$$

Measures of Variability or Dispersion (Spread)

- Quantify how much the **data values in a dataset differ from the central value** (mean or median).

- These measures help **describe the distribution's spread** and indicate how concentrated or scattered the data is.

Range	○ Difference between maximum and minimum values
Interquartile Range	○ Difference between third and first quartile (Q3 - Q1)
Variance	○ Average of the squared deviations from the mean ○ Definitions of population variance and sample variance differ slightly
Standard Deviation	○ Square root of the variance

Example – Range and Interquartile Range

Sales	Sorted Sales	Rank	
9	6	1	Minimum
6	9	2	
12	10	3	
10	12	4	
13	13	5	First Quartile
15	14	6	
16	14	7	
14	15	8	
14	16	9	
16	16	10	
17	16	11	
16	17	12	
24	17	13	
21	18	14	
22	18	15	Third Quartile
18	19	16	
19	20	17	
18	21	18	
20	22	19	
17	24	20	Maximum

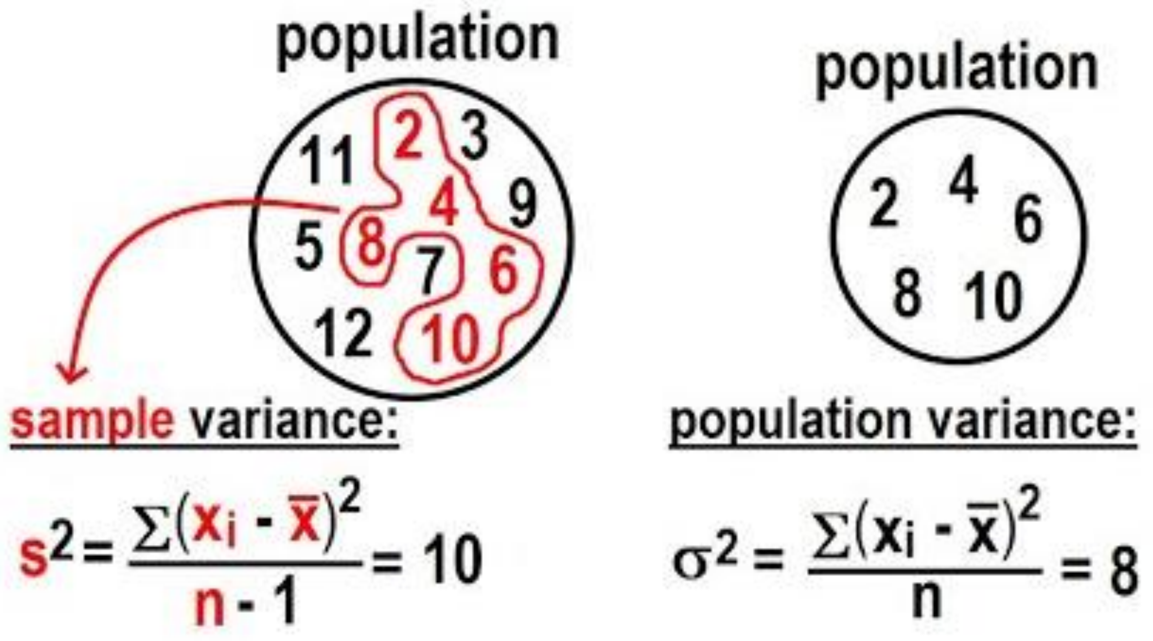
Range: Maximum - Minimum =
24 - 6 = 18

Q_1 = 13 + (.25)(1) = 13.25

Q_3 = 18+ (.75)(1) = 18.75

Interquartile Range: Q3 - Q1 =
18.75 - 13.25 = 5.5

Variance and Standard Deviation



sample variance:

$$s^2 = \frac{\Sigma(x_i - \overline{x})^2}{n - 1} = 10$$

population variance:

$$\sigma^2 = \frac{\Sigma(x_i - \overline{x})^2}{n} = 8$$

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\overline{x}$ = sample mean n = sample size

Calculation of Sample Variance

x	$x-\bar{x}$	$(x-\bar{x})^2$	x^2
6	-9.85	97.0225 36	
9	-6.85	46.9225 81	
10	-5.85	34.2225 100	
12	-3.85	14.8225 144	
13	-2.85	8.1225	169
14	-1.85	3.4225	196
14	-1.85	3.4225 196	
15	-0.85	0.7225 225	
16	0.15	0.0225 256	
16	0.15	0.0225 256	
16	0.15	0.0225 256	
17	1.15	1.3225 289	
17	1.15	1.3225 289	
18	2.15	4.6225 324	
18	2.15	4.6225 324	
19	3.15	9.9225 361	
20	4.15	17.2225 400	
21	5.15	26.5225 441	
22	6.15	37.8225 484	
24	8.15	66.4225 576	
317	0	378.5500	5403

$$s^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{(n-1)} = \frac{378.55}{(20-1)}$$

$$= \frac{378.55}{19} = 19.923684$$

$$= \frac{\sum_{i=1}^{n} x^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}}{(n-1)}$$

$$= \frac{5403 - \frac{317^2}{20}}{(20-1)} = \frac{5403 - \frac{100489}{20}}{19}$$

$$= \frac{5403-5024.45}{19} = \frac{378.55}{19} = 19.923684$$

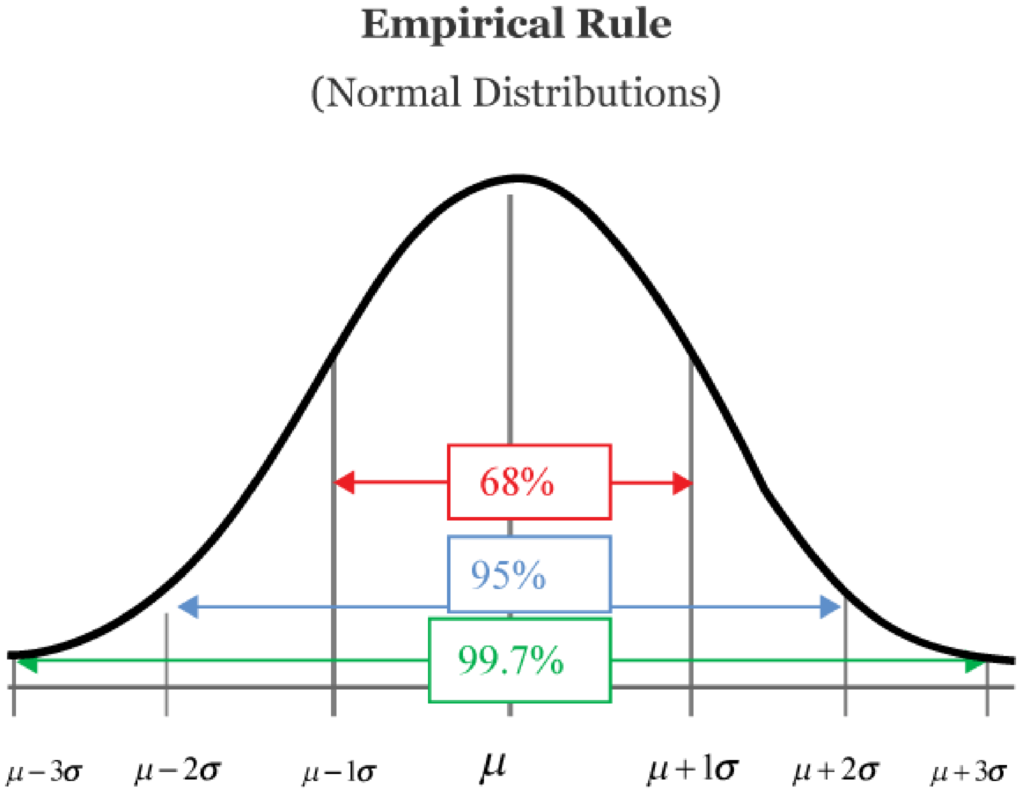
$$s = \sqrt{s^2} = \sqrt{19.923684} = 4.46$$

Relations between the Mean and Standard Deviation

- The **mean** represents the average value of a dataset or the central point around which the data is distributed.
- The **standard deviation** measures the spread or dispersion of the data points from the mean. It quantifies how far, on average, the data points are from the mean.
- A **low standard deviation** means that the data points are close to the mean (the data is less spread out).
- A **high standard deviation** means that the data points are more spread out from the mean (the data is more dispersed).

Relations between the Mean and Standard Deviation

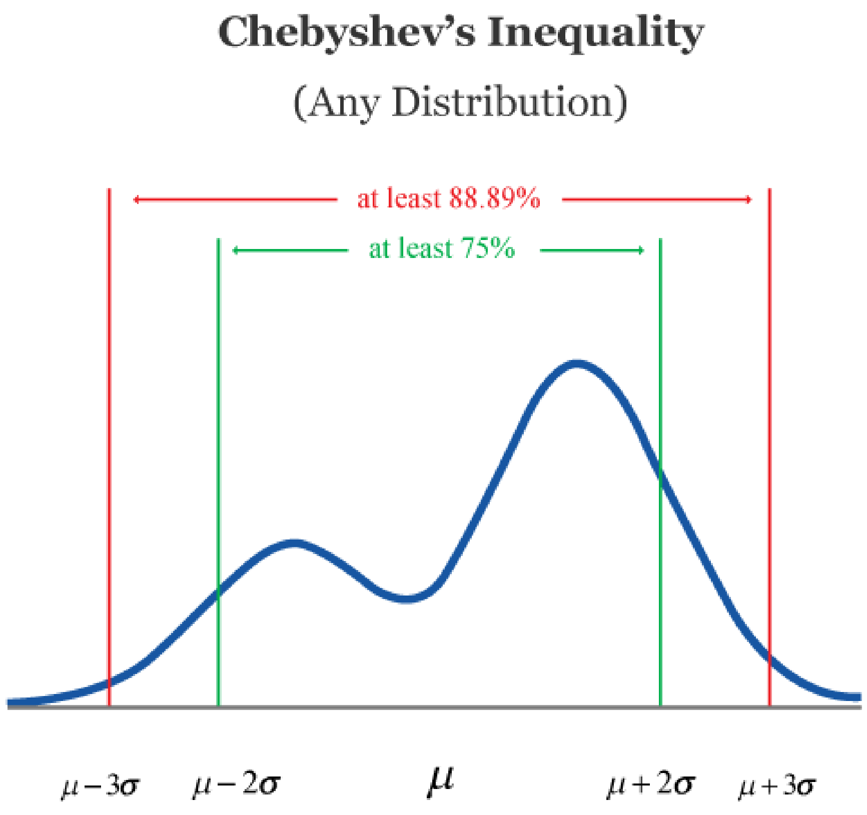
- **Empirical Rule**
  - In a normal distribution (bell-shaped curve), about **68%** of the data points lie within one standard deviation of the mean, **95%** within two standard deviations, and **99.7%** within three standard deviations.
  - This is known as the **68 – 95 – 99.7** rule or the empirical rule.



Relations between the Mean and Standard Deviation

Assignments

Pelajari tentang Chebyshev’s Inequality ☺



Chebyshev’s Inequality

(Any Distribution)

Exercise 1

- Jelaskan dan pelajari tentang **Aturan Empiris (Empirical Rules)**

- Kerjakan:
  - Tinggi badan siswa di sebuah sekolah mengikuti distribusi normal dengan rata-rata 160 cm dan standar deviasi 7 cm. Gunakan Aturan Empiris untuk menjawab pertanyaan berikut:
    - Berapa rentang tinggi badan di mana sekitar 68% siswa berada?
    - Berapa rentang tinggi badan di mana sekitar 95% siswa berada?
    - Berapa rentang tinggi badan di mana sekitar 99.7% siswa berada?

Exercise 2

- Jelaskan dan pelajari tentang **Aturan Empiris (Empirical Rules)**

- Kerjakan:
  - Dalam sebuah uji coba, waktu reaksi dari sejumlah pengemudi diukur. Diketahui bahwa waktu reaksi rata-rata adalah 0,8 detik dengan standar deviasi 0,1 detik. Berdasarkan Aturan Empiris, tentukan rentang waktu reaksi di mana:
    - 68% pengemudi berada
    - 95% pengemudi berada
    - 99.7% pengemudi berada

Exercise 3

- Jelaskan dan pelajari tentang **Teorema Chebyshev (Chebyshev’s Theorem)**

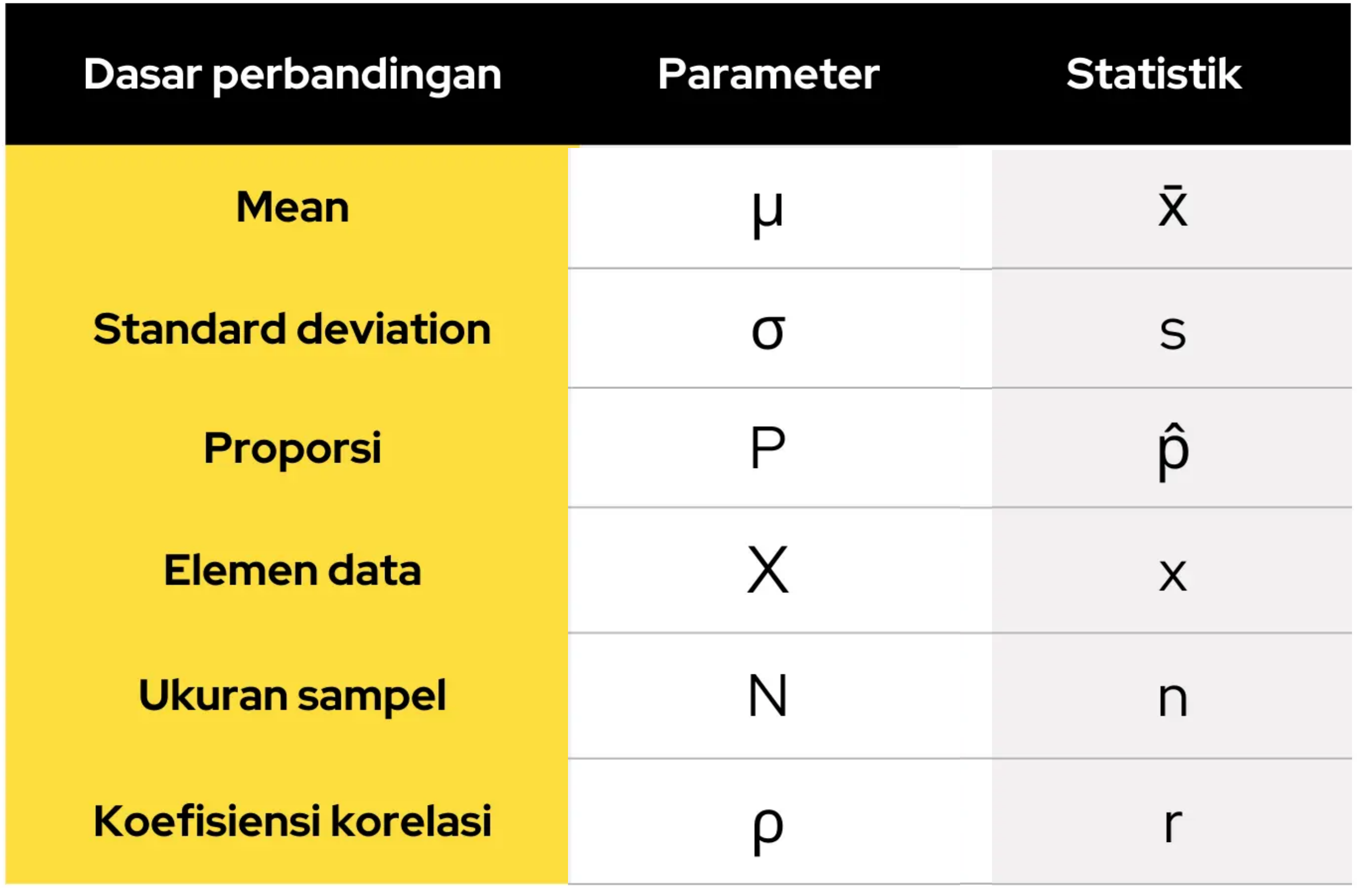
- Kerjakan:
  - Sebuah perusahaan mencatat waktu produksi barang dengan rata-rata 40 menit dan standar deviasi 5 menit. Gunakan Teorema Chebyshev untuk menjawab pertanyaan berikut:
    - Berapa proporsi minimum waktu produksi yang berada dalam jarak 3 standar deviasi dari rata-rata?

Exercise 4

- Jelaskan dan pelajari tentang **Teorema Chebyshev (Chebyshev's Theorem)**

- Kerjakan:
  - Dari sebuah penelitian, diketahui bahwa penghasilan bulanan dari 100 orang karyawan memiliki rata-rata Rp5.000.000 dengan standar deviasi Rp500.000. Tentukan proporsi minimum dari karyawan yang penghasilannya berada dalam jarak 2 standar deviasi dari rata-rata menurut **Teorema Chebyshev**.

Notation

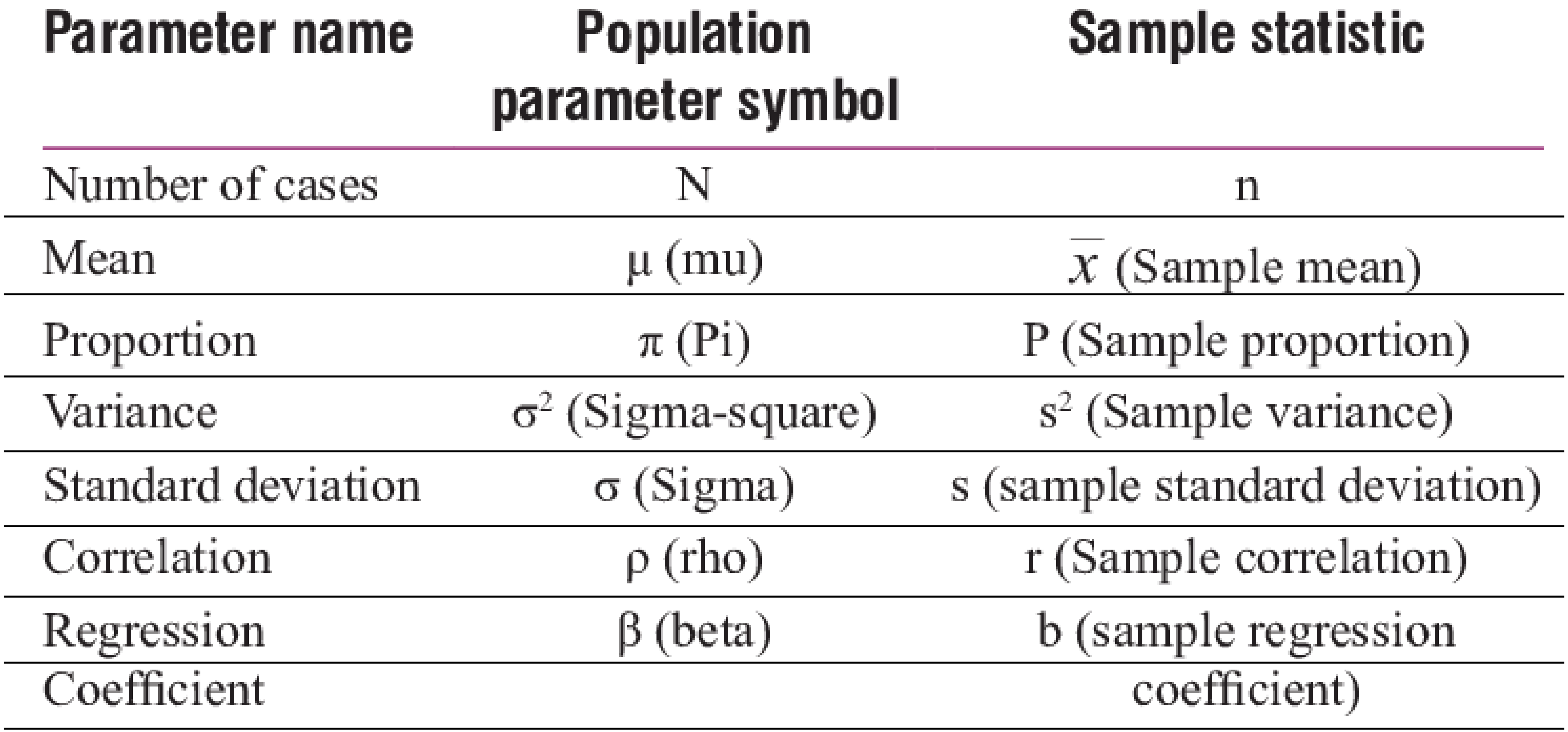


Dasar perbandingan	Parameter	Statistik
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proporsi	P	$\hat{p}$
Elemen data	X	x
Ukuran sampel	N	n
Koefisiensi korelasi	ρ	r

Sumber: Key Differences

https://revou.co/revoupedia/kosakata

Notation



Parameter name	Population parameter symbol	Sample statistic
Number of cases	N	n
Mean	μ (mu)	$\overline{x}$ (Sample mean)
Proportion	π (Pi)	P (Sample proportion)
Variance	σ^2 (Sigma-square)	s^2 (Sample variance)
Standard deviation	σ (Sigma)	s (sample standard deviation)
Correlation	ρ (rho)	r (Sample correlation)
Regression Coefficient	β (beta)	b (sample regression coefficient)