Applied Statistical Modeling And Data Analytics

Introduction to Statistics

Assistant Professor Dr. Pornpimol Chaiwuttisak

Module 05177003 : Applied Statistical Modeling and Data Analytics

เนื้อหา	อาจารย์ผู้รับผิดชอบ	คะแนน(%)
ความรู้เบื้องต้นเกี่ยวสถิติ ความน่าจะเป็น การทดสอบไคสแควร์	ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์	20
การวิเคราะห์การถดถอยเชิงเส้นอย่างง่าย การวิเคราะห์การถดถอยพหุคูณ การถดถอยโลจิสติค การวิเคราะห์การถดถอยโลจิสติคพหุคูณ	ดร.ยุวดี กล่อมวิเศษ	30
การจำแนกข้อมูล การจัดกลุ่มของข้อมูล ตัวอย่างกรณีศึกษาจากโจทย์ภาคธุรกิจและภาคอุตสาหกรรม	รศ.สายชล สินสมบูรณ์ทอง	50

คณาจารย์

ผศ.ดร.พรพิมล ชัยวุฒิศักดิ์

ดร.ยุวดี กล่อมวิเศษ

รศ.สายชล สินสมบูรณ์ทอง

Assistant Professor Dr. Pornpimol Chaiwuttisak

- **Ph.D. (Mathematical Science)**
  Faculty of Mathematical Science
  University of Southampton, United Kingdom

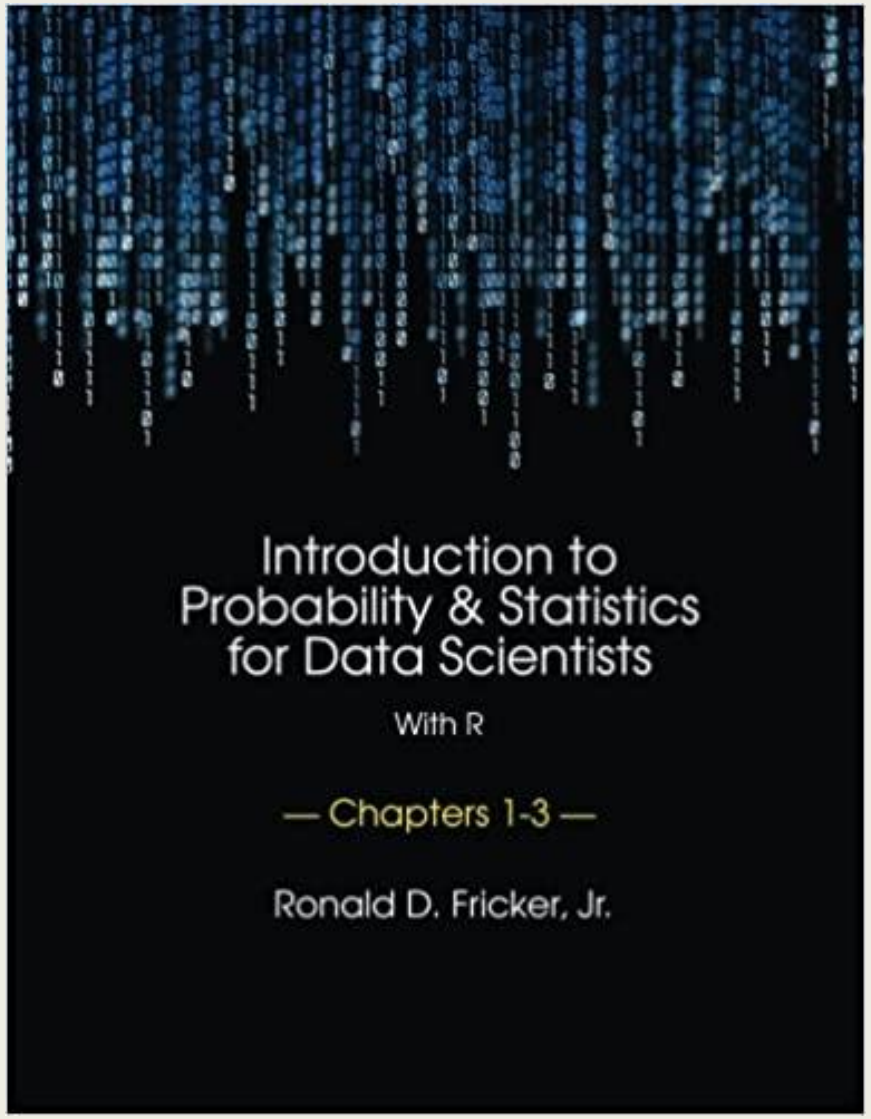
- **M.Sc. (Operational Research)**
  Faculty of Mathematical Science
  University of Southampton, United Kingdom

- **M.Sc. (Applied Statistics)**
  National Institute of Development Administration (NIDA), Bangkok
  THAILAND

- **B.Sc. (Computer Science)**
  Kasetsart University, Bangkok THAILAND

Textbooks

- Hanke E. J, Reitsch A. G: Understanding Business Statistics
- Anderson, D.R. - Sweeney, D.J. - Williams, T.A.: Statistics for Business and Economics. South-Western Pub., 2005, 320 p., ISBN 978-032-422-486-3
- Jaisingh, L.R.: Statistics for the Utterly Confused. McGraw Hill, 2005, 352 p., ISBN 978-007-146-193-1
- Everitt, B. S.: The Cambridge (explanatory) dictionary of statistics. Cambridge University Press, 2006, 442 p., ISBN 978-052-169-027-0
- Illowsky, B. - Dean, S. (2009, August 5). Collaborative Statistics. Retrieved from the Connexions Web site: http://cnx.org/content/col10522/1.36



Material and Assignment

Team Code : uijt0an

Group Line

Scan QR Code

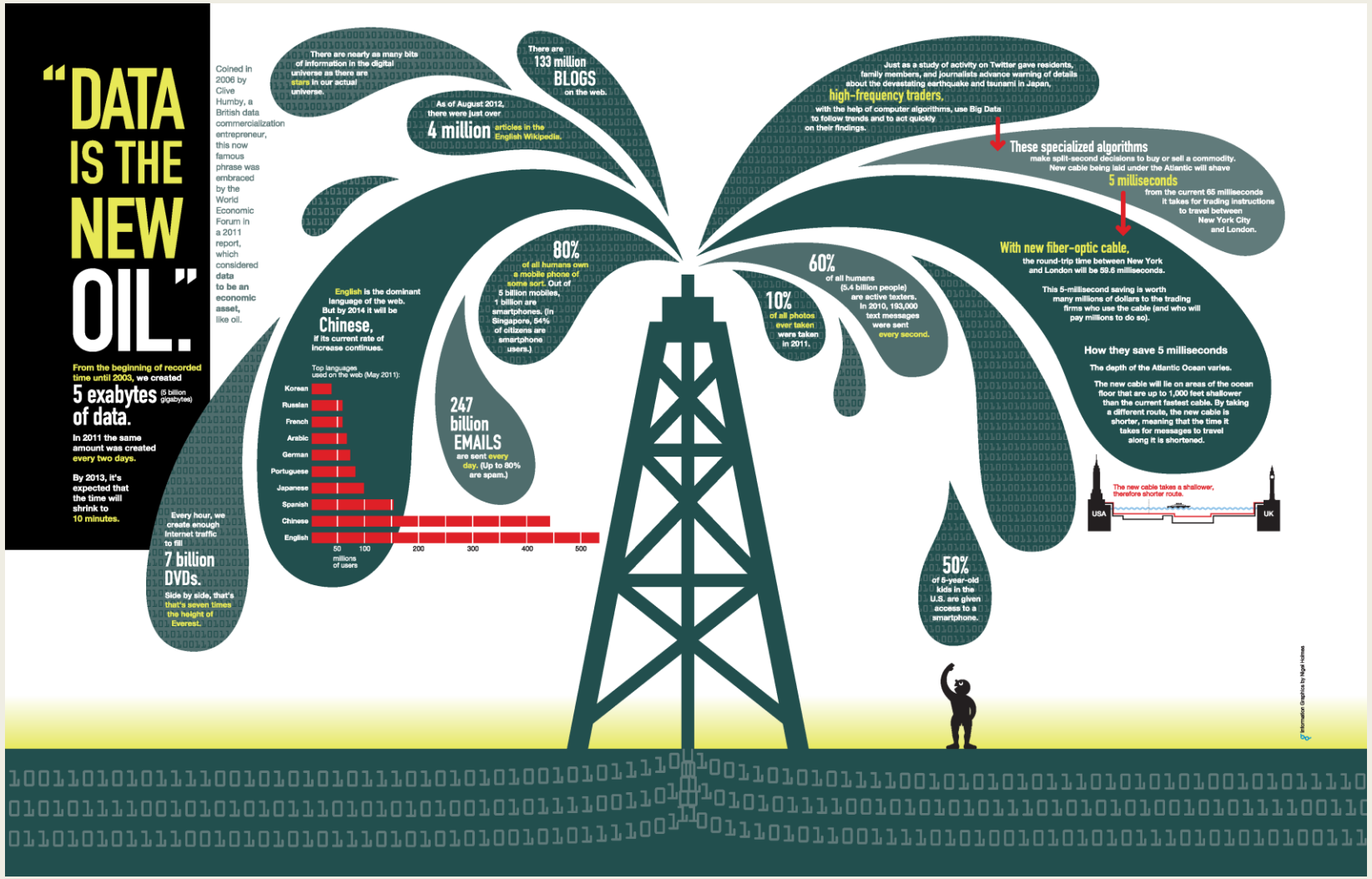
Vote: http://etc.ch/kPcn

Cockpit: https://directpoll.com/c?XDVhEtjtb9A12Cm694bANl6XCe7P1Ir

Pre-test

shorturl.at/kpv29

"Data is the New Oil" – World Economic Forum 2011



คำกล่าวของ John Wanamaker

“Half the money I spend on advertising is wasted; the trouble is I don’t know which half.”

การกระทำหรือการตัดสินใจอะไรก็ตามที่ไม่ได้มีการใช้ข้อมูลย่อมทำให้เกิดความสูญเปล่าได้

คำกล่าวของ Seth Godin

"Data is not useful until it becomes information"

ข้อมูลไม่มีประโยชน์จนกลายเป็นข่าวสารหรือสารสนเทศ

"Data Science" an Emerging Field

O'Reilly Radar report, 2011

Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Goal of Data Science

Turn data into data products.

How to use data?

- Data => exploratory analysis => knowledge models => product / decision marking
- Data => predictive models => evaluate / interpret => product / decision making

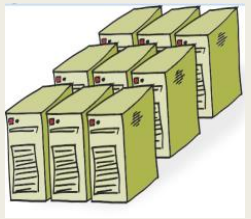
Data Scientist's Practice

Digging Around in Data

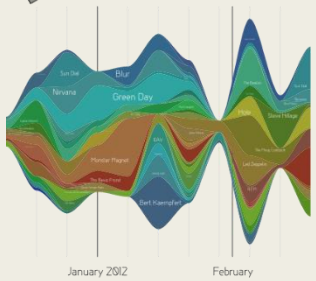
Clean, prep

$$\begin{bmatrix} \cos 90^\circ & \sin 90^\circ \\ -\sin 90^\circ & \cos 90^\circ \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \vec{a} \\ \vec{a} \end{bmatrix}$$

Hypothesize Model



Large Scale Exploitation



Evaluate Interpret

Example data science applications

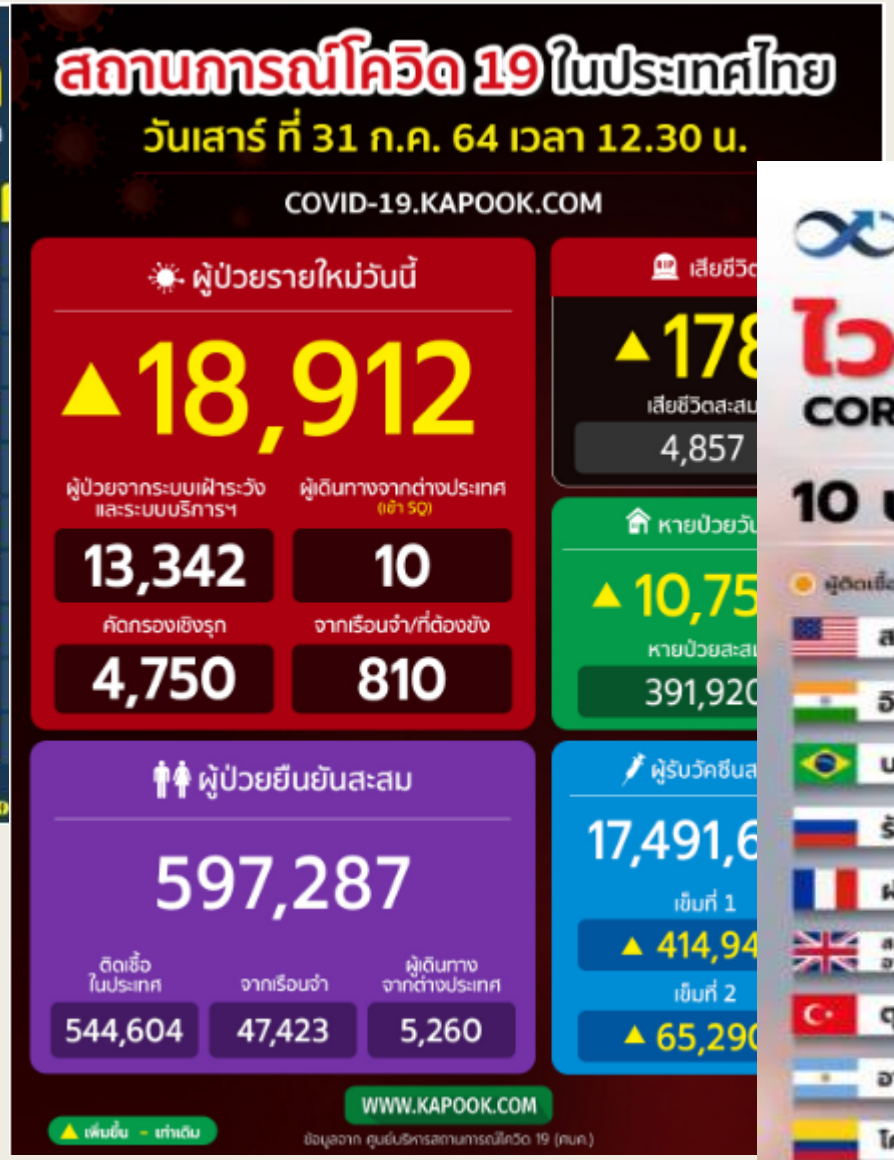
- Marketing: predict the characteristics of high life time value (LTV) customers, which can be used to support customer segmentation, identify upsell opportunities, and support other marking initiatives
- Logistics: forecast how many of which things you need and where will we need them, which enables learn inventory and prevents out of stock situations
- Healthcare: analyze survival statistics for different patient attributes (age, blood type, gender, etc.) and treatments; predict risk of re-admittance based on patient attributes, medical history, etc.

More Examples

- Transaction Databases → Recommender systems (NetFlix), Fraud Detection (Security and Privacy)
- Wireless Sensor Data → Smart Home, Real-time Monitoring, Internet of Things
- Text Data, Social Media Data → Product Review and Consumer Satisfaction (Facebook, Twitter, LinkedIn), E-discovery
- Software Log Data → Automatic Trouble Shooting (Splunk)
- Genotype and Phenotype Data → Epic, 23andme, Patient-Centered Care, Personalized Medicine

Statistics

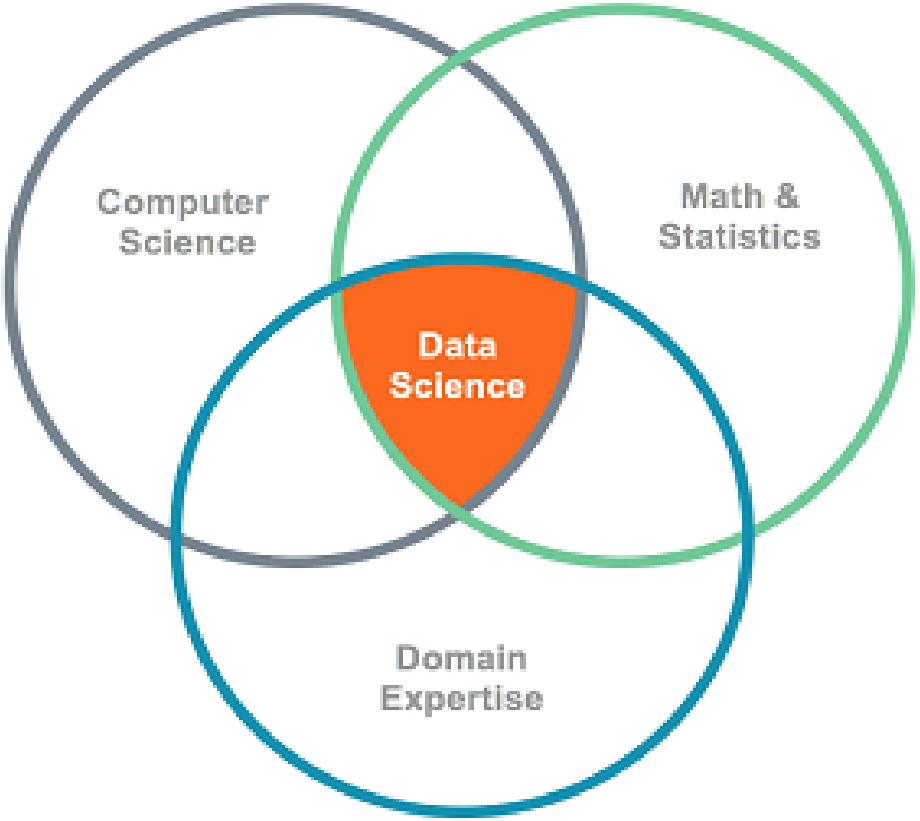

10 จังหวัด
ติดเชื้อโควิด-19 สูงสุด
วันนี้
1 กรุงเทพมหานคร 3,231
2 สมุทรปราการ 1,386
3 สมุทรสาคร 1,186
4 ชลบุรี 914
5 นนทบุรี 587
6 ฉะเชิงเทรา 479
7 นครปฐม 378
8 อุบลราชธานี 350
9 ปทุมธานี 330
10 สงขลา 324


สถานการณ์โควิด 19 ในประเทศไทย
วันเสาร์ ที่ 31 ก.ค. 64 เวลา 12.30 น.
COVID-19.KAPOOK.COM
ผู้ป่วยรายใหม่วันนี้
▲18,912
ผู้ป่วยจากระบบเฝ้าระวังและระบบบริการฯ
13,342
ผู้เดินทางจากต่างประเทศ (เข้า SQ)
10
คัดกรองเชิงรุก
4,750
จากเรือนจำ/ที่ต้องขัง
810
ผู้ป่วยยืนยันสะสม
597,287
ติดเชื้อในประเทศ
544,604
จากเรือนจำ
47,423
ผู้เดินทางจากต่างประเทศ
5,260
เสียชีวิต
▲178
เสียชีวิตสะสม
4,857
หายป่วยวัน
▲10,75
หายป่วยสะสม
391,920
ผู้รับวัคซีน
17,491,6
เข็มที่ 1
▲414,94
เข็มที่ 2
▲65,290
WWW.KAPOOK.COM
เพิ่มขึ้น - เท่าเดิม


THE BANGKOK INSIGHT
อัพเดท ณ เวลา 09.00 น.
ไวรัสโคโรนา
CORONAVIRUS COVID-19
กรกฎาคม 31
10 ประเทศติดเชื้อสูงสุดทั่วโลก
ผู้ติดเชื้อรวม ผู้ติดเชื้อรายใหม่ ผู้เสียชีวิตรวม รักษาหายแล้วรวม
สหรัฐ 35,688,506 99,470 629,064 29,652,038
อินเดีย 31,612,794 41,499 423,842 30,773,555
บราซิล 19,880,273 40,904 555,512 18,595,380
รัสเซีย 6,242,066 23,564 157,771 5,588,848
ฝรั่งเศส 6,103,548 24,309 111,824 5,696,619
สหราชอาณาจักร 5,830,774 29,622 129,583 4,498,089
ตุรกี 5,704,713 22,083 51,253 5,449,253
อาร์เจนตินา 4,919,408 13,483 105,586 4,557,037
โคลอมเบีย 4,776,291 9,462 120,432 4,567,701
สเปน 4,447,044 24,753 81,486 3,711,200
www.worldometers.info/coronavirus/ | www.thebangkokinsight.com
COVID-19

Data Science VS Statistics

Data science is a powerful combination of various disciplines.



Computer Science Skills

- Programming
- Big data technologies

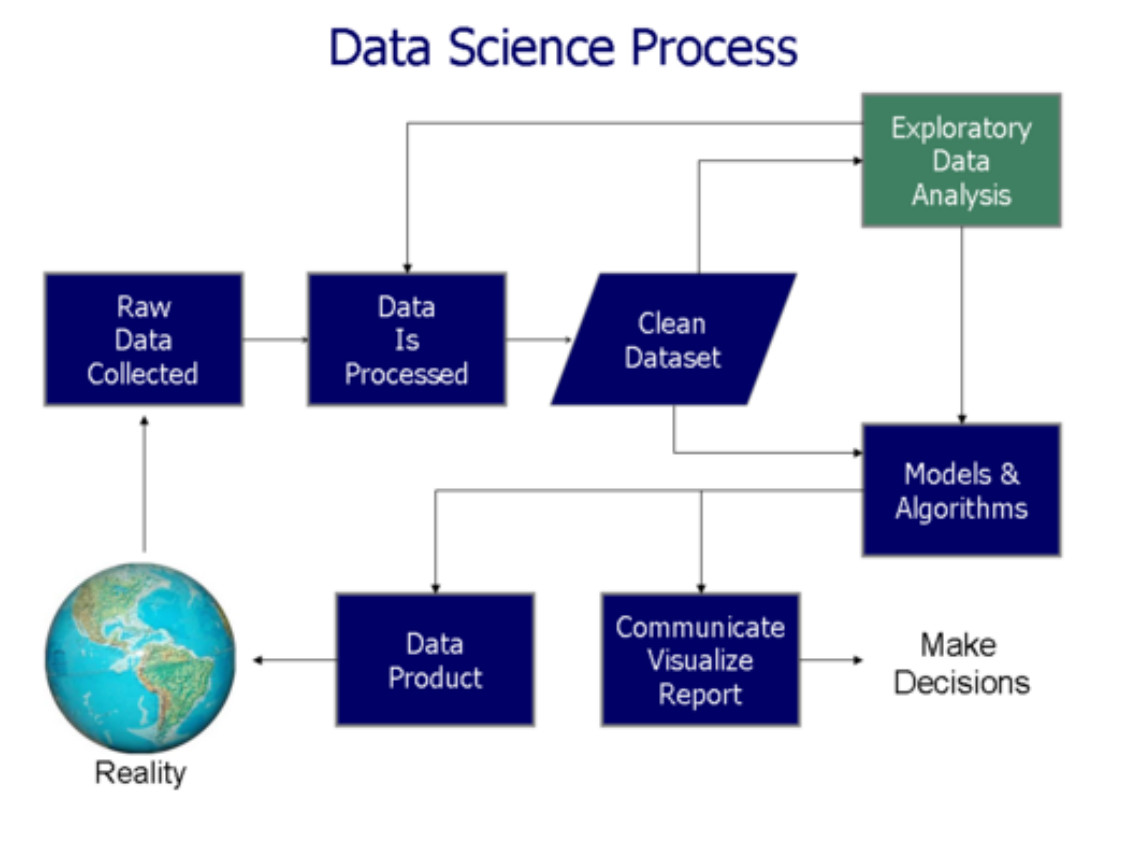
Math and Statistics Knowledge

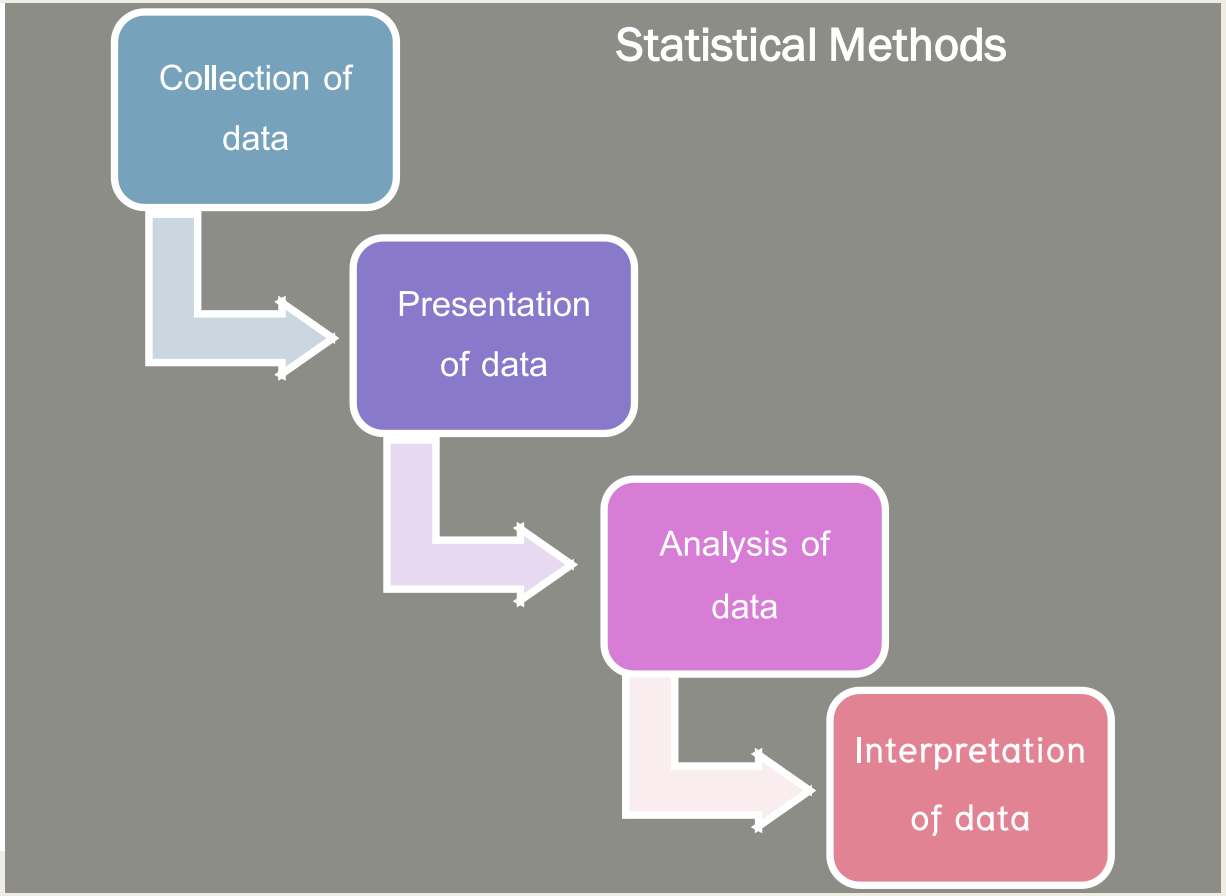
- Machine learning
- Ensemble models
- Anomaly detection

Domain Expertise

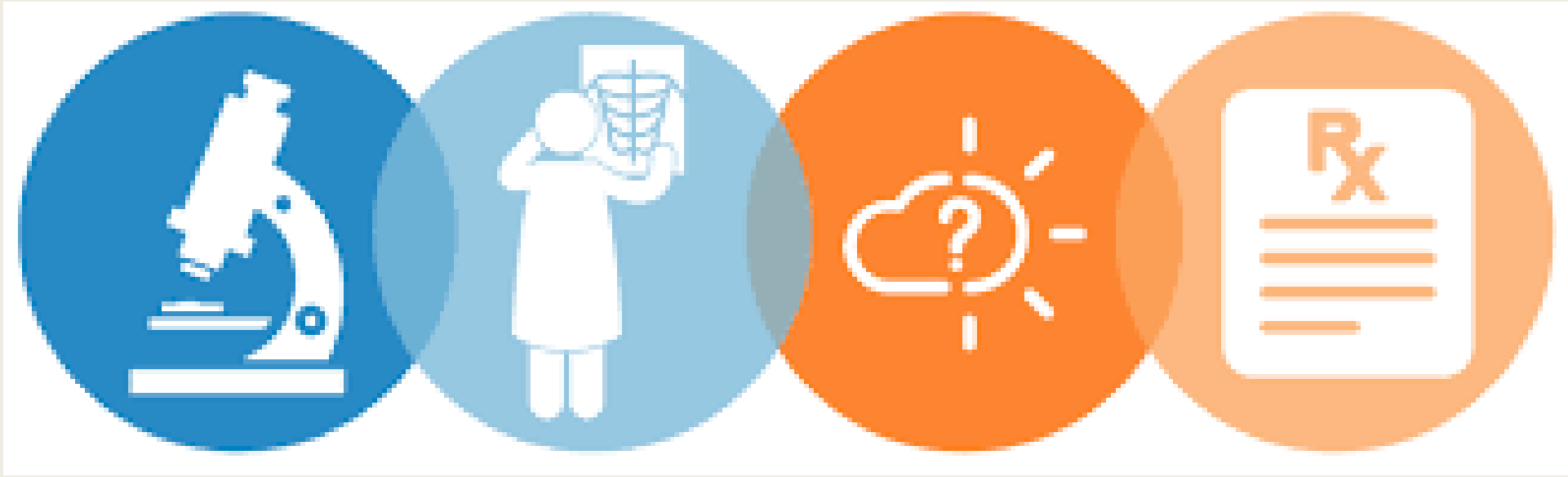
- Business knowledge
- Expert systems
- User testing

Data Science VS Statistics





Data Analytical Level



Descriptive	Diagnostic	Predictive	Prescriptive
Explains what happened	Explains why it happened	Forecasts what might happen	Recommends an action based on the forecast

Statistics and Statistic

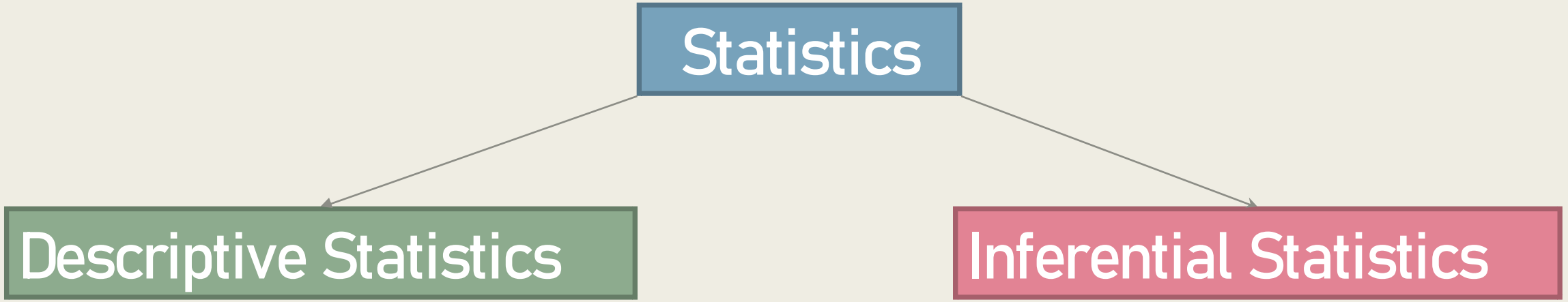
a) **Statistics** is the course you are studying right now, also known as **statistical analysis**. It is a field of study concerned with summarizing data, interpreting data, making decisions based on data

b) A quantity calculated in a **sample** to estimate a value in a population is called a "**statistic**"

- The related term **data science** or **data analysis** stands for a study of processes and systems that extract knowledge or insights from data in various forms, either structured or unstructured.

- Data Science is a continuation of some of the fields such as statistics, data mining, and predictive analytics

Structure of Statistics



Presenting, organizing and summarizing data

Drawing conclusions about a population based on data observed in a sample

Data analysis processes

Data collection and preparation

- Collect data
- Prepare codebook
- Setup structure
- Enter data
- Screen data for errors

Exploration of data

- Summary table
- Graphs

Analysis

- Explore relationship between variables
- Compare groups

Data analysis and Descriptive statistics

Data collection and preparation	Exploration of data	Analysis
Collect data	Summary table	Explore relationship between variables
Prepare codebook		
Setup structure	Graphs	Compare groups
Enter data		
Screen data for errors		

Descriptive statistics

Data analysis and Inferential statistics

Data collection and preparation

- Collect data
- Prepare codebook
- Setup structure
- Enter data
- Screen data for errors

Exploration of data

- Summary table
- Graphs

Analysis

- Explore relationship between variables
- Compare groups

Inferential statistics

Descriptive Statistics (DS)

- Descriptive statistics is used to **summarize** and **describe data**
- Descriptive statistics is a collection of methods for summarizing data such as mean, median, mode, range, variance

An example of DS: Salaries in Estonia

WS720: AVERAGE GROSS HOURLY EARNINGS OF FULL-TIME AND PART-TIME EMPLOYEES, OCTOBER by Year, Major group of occupation and Age group

	Less than 30	30–39	40–49	50–59	60 and over
2008					
Total	5.01	6.04	5.30	4.76	4.09
Legislators, senior officials and managers	7.11	8.59	7.38	7.05	6.21
Professionals	7.10	7.96	7.45	6.67	6.06
Technicians and associate professionals	5.46	6.19	5.52	4.82	4.27
Clerks	4.49	5.09	4.30	4.11	3.54
Service workers and shop and market sales workers	3.70	3.93	3.69	3.23	2.63
Skilled agricultural and fishery workers	3.20	3.46	3.52	3.56	2.80
Craft and related trades workers	4.64	5.01	4.71	4.40	3.79
Plant and machine operators and assemblers	4.38	4.68	4.28	4.02	3.74
Elementary occupations	3.70	3.49	3.07	2.77	2.21
Armed forces	..	..	..	..	..

Footnote:
Unit: euros
The data have been converted into euros on the basis of aggregated data (1 euro = 15.6466 Estonian kroons).
The data are in compliance with ISCO 88. Data are continued to be published according to the new classification in table WS640: Average gross hourly earnings of full-time and part-time employees by major group of occupation, sex and age group, October.

Example of DS: Winners of Olympic marathon

	WOMEN		
Year	Winner	Country	Time
1984	Joan Benoit	USA	2:24:52
1988	Rosa Mota	POR	2:25:40
1992	Valentina Yegorova	UT	2:32:41
1996	Fatuma Roba	ETH	2:26:05"
2000	Naoko Takahashi	JPN	2:23:14
2004	Mizuki Noguchi	JPN	2:26:20
2008	Constantina Tomescu	ROU	2:26:44
2012	Tiki Gelana	ETH	2:23:07

	MEN						
Year	Winner	Country	Time	Year	Winner	Country	Time
1896	Spiridon Louis	GRE	2:58:50	1960	Abebe Bikila	ETH	2:15:16
1900	Michel Theato	FRA	2:59:45	1964	Abebe Bikila	ETH	2:12:11
1904	Thomas Hicks	USA	3:28:53	1968	Mamo Wolde	ETH	2:20:26
1906	Billy Sherring	CAN	2:51:23	1972	Frank Shorter	USA	2:12:19
1908	Johnny Hayes	USA	2:55:18	1976	Waldemar Cierpinski	E.Ger	2:09:55
1912	Kenneth McArthur	S. Afr.	2:36:54	1980	Waldemar Cierpinski	E.Ger	2:11:03
1920	Hannes Kolehmainen	FIN	2:32:35	1984	Carlos Lopes	POR	2:09:21
1924	Albin Stenroos	FIN	2:41:22	1988	Gelindo Bordin	ITA	2:10:32
1928	Boughra El Ouafi	FRA	2:32:57	1992	Hwang Young-Cho	S. Kor	2:13:23
1932	Juan Carlos Zabala	ARG	2:31:36	1996	Josia Thugwane	S. Afr.	2:12:36
1936	Sohn Kee-Chung	JPN	2:29:19	2000	Gezahenge Abera	ETH	2:10:10
1948	Delfo Cabrera	ARG	2:34:51	2004	Stefano Baldini	ITA	2:10:55
1952	Emil Ztopek	CZE	2:23:03	2008	Samuel Wanjiru	KEN	2:06:32
1956	Alain Mimoun	FRA	2:25:00	2012	Stephen Kiprotich	UGA	2:08:01

Inferential Statistics

- **Estimation**
  - e.g., Estimate the population mean weight using the sample mean weight
- **Hypothesis testing**
  - e.g., Test the claim that the population mean weight is 70 kg

Inference is the process of drawing conclusions or making decisions about a population based on sample results

Populations and Samples

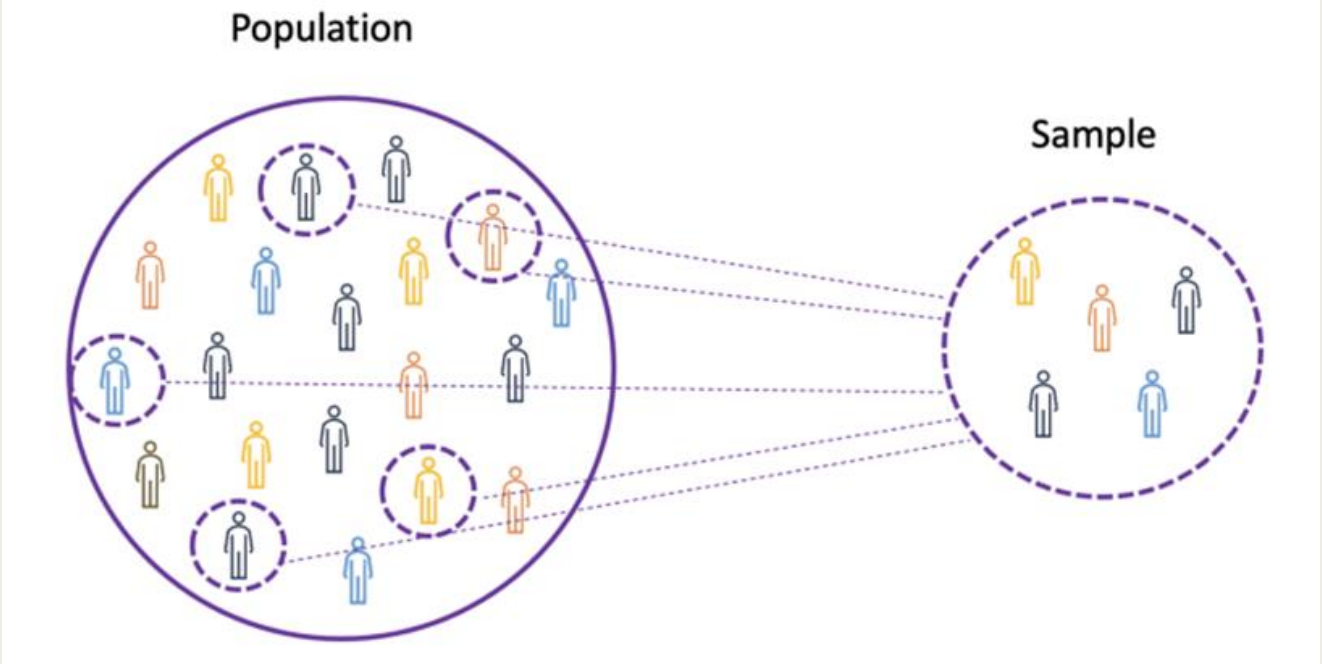
Gathering all data is not always possible due to barriers such as time, accessibility, or cost instead of that, we often gather information from a smaller subset of the population, known as a sample

- Population: The entire set of possible observations in which we are interested
- Sample: A subset of the population from which information is actually collected

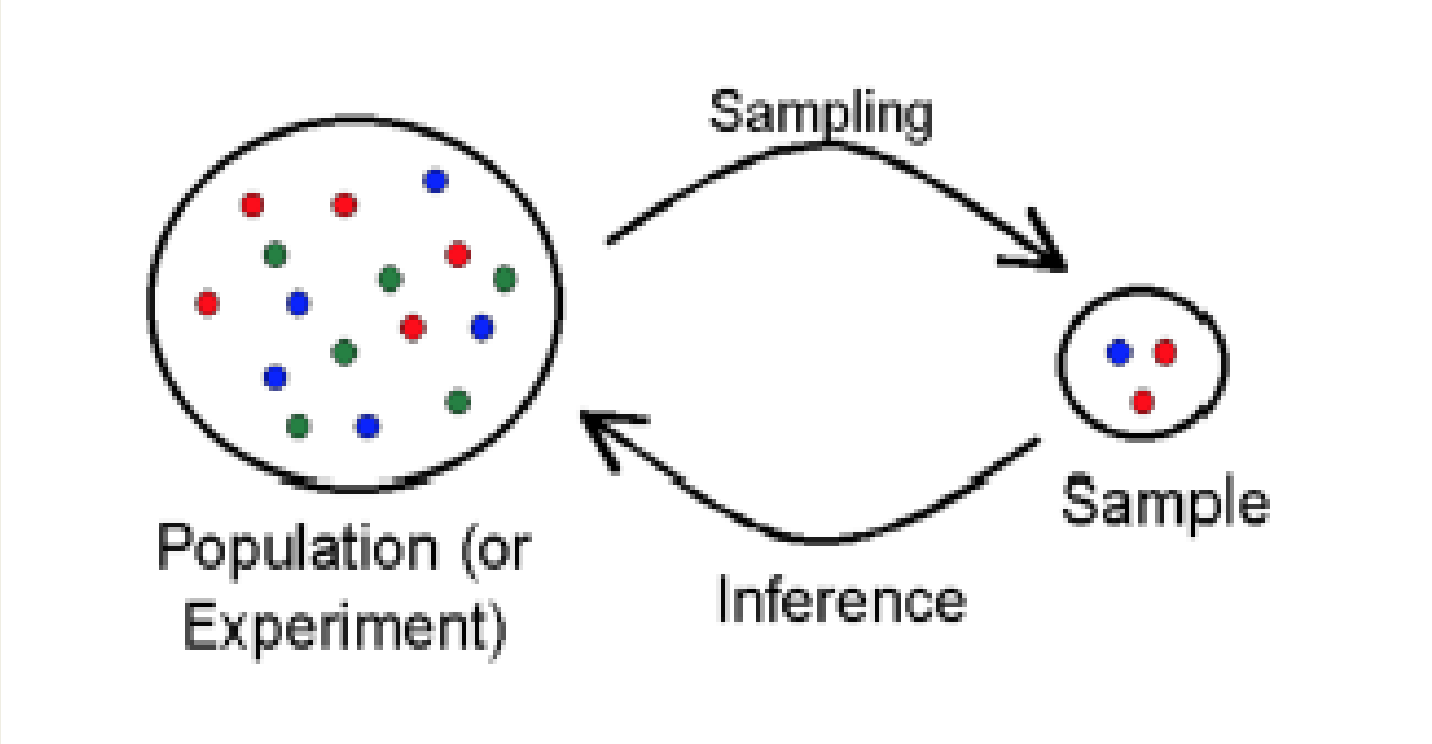
Inferential statistics is a collection of methods for using sample data to make conclusion about a population

Populations and Samples

- Usually use ***N*** to denote the total number of observations in population.
- Usually use ***n*** to denote total number of observations in sample.



Populations and Samples



Terminology and Notation

- **Populations** have parameters. A descriptive measure of a population that is usually unobservable and unknown.
- **Parameters** are usually denoted by **Greek letters**:
  - Population average/mean: μ
  - Population variance: σ^2

Terminology and Notation

- **Samples** have statistic: A descriptive measure of a sample that can be observed (calculated) and is known.
- **Sample statistic** is used to make inferences about population parameters and are usually denoted by **Roman letters**:
  - Sample average/mean: $\bar{x}$
  - Sample variance: s^2

Population and Sample

- Given that **experimenting with an entire population is** either **impossible or simply too expensive**, researchers or analysts use samples rather than the entire population in their experiments or trials.
- To make sure that the experimental results are reliable and hold for the entire population, **the sample needs to be a true representation of the population.** That is, the sample needs to be unbiased.

.

Kinds of sample

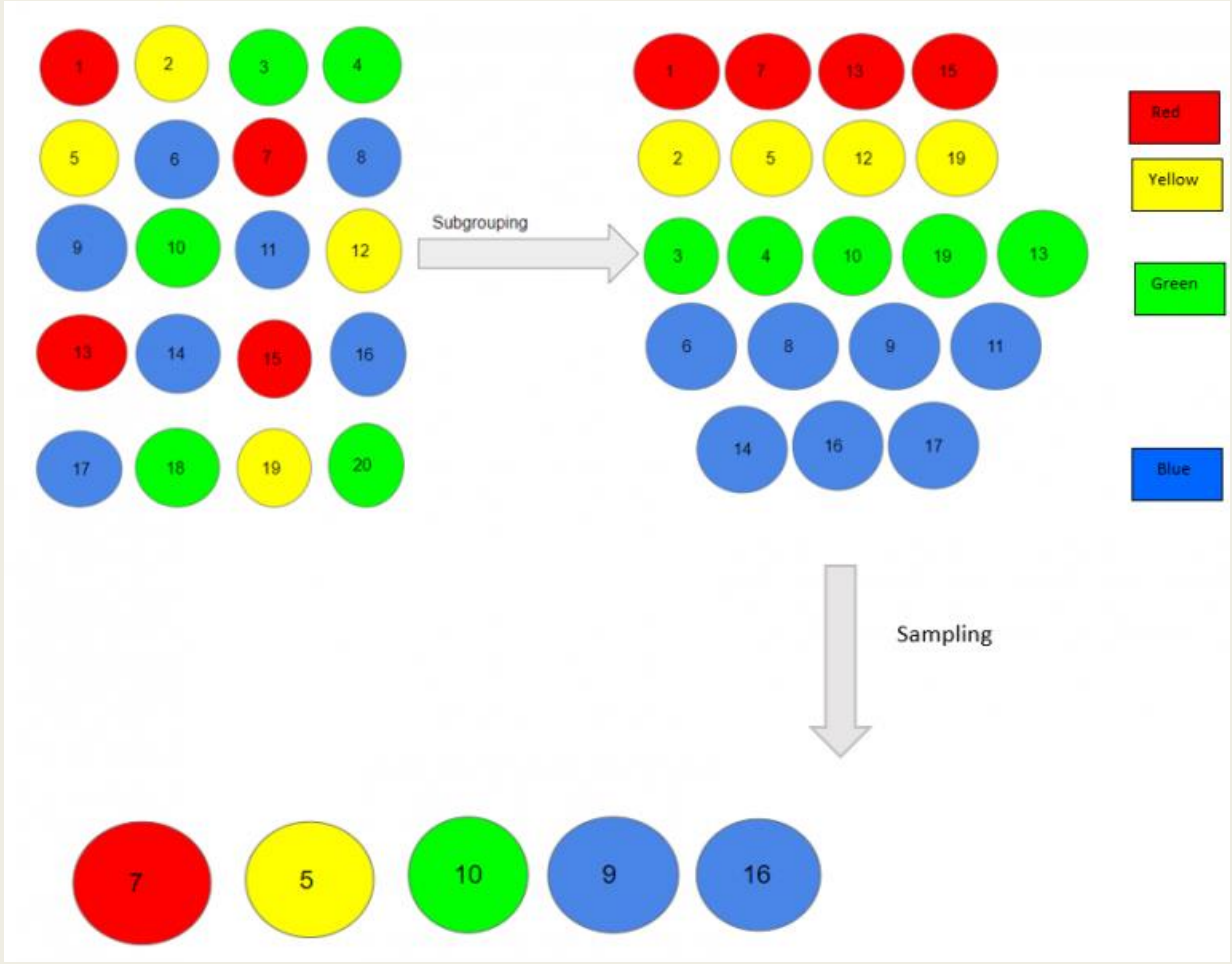
- A compete sample is a set of objects from a parent population that includes ALL such objects that satisfy a set of well-defined selection criteria
- **An unbiased (representative) sample** is a set of objects chosen from a complete sample using a sampling method which is free from bias.

Sampling methods

- **Simple Random Sampling (SRS)**: Every member and set of members has an equal chance of being included in the sample.

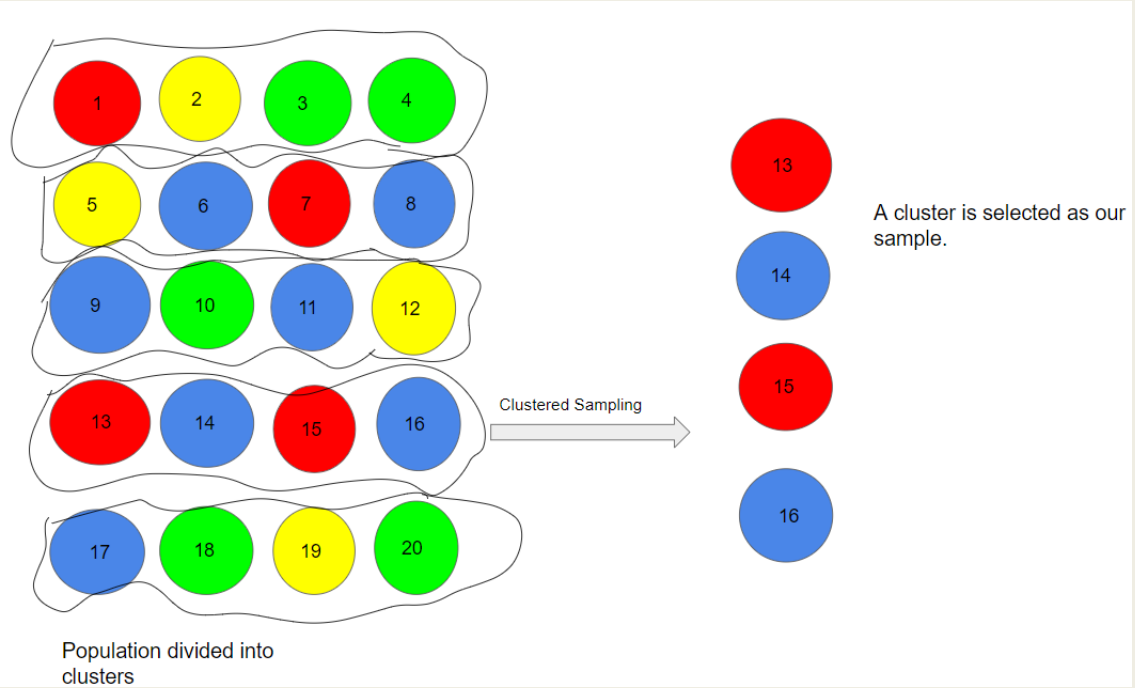
Sampling Methods

Stratified Random Sampling: The population is first split into groups (Strata). The overall sample consists of some members from every group. The members from each group are chosen randomly.



Sampling Methods

- **Cluster Random Sampling:** The population is first split into groups (Cluster). The overall sample consists of every member from some of the groups. The groups are selected at random

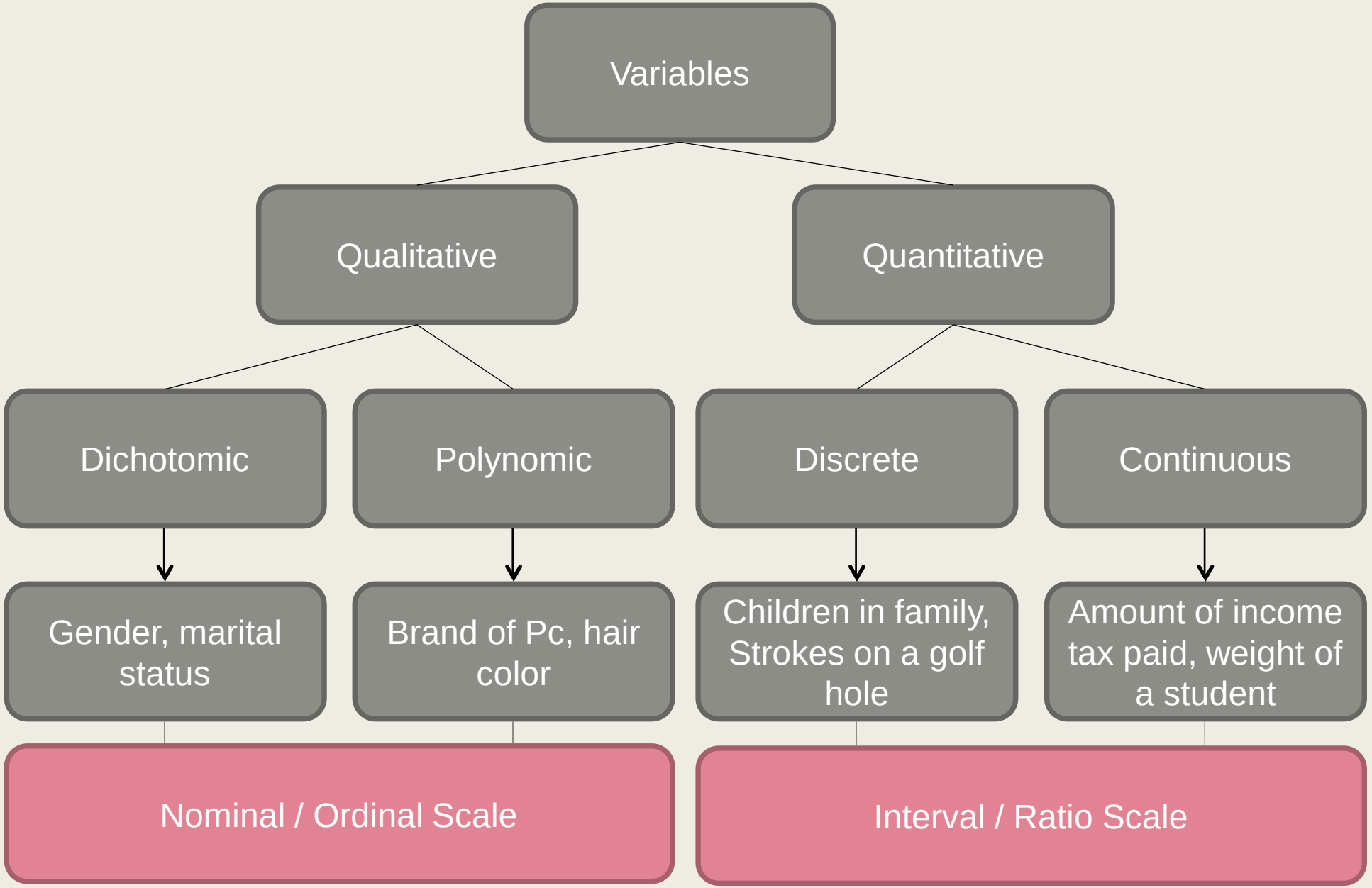


- Example: An airline company wants to survey its customers one day, so they randomly select 5 flights that day and survey every passenger on those flights.

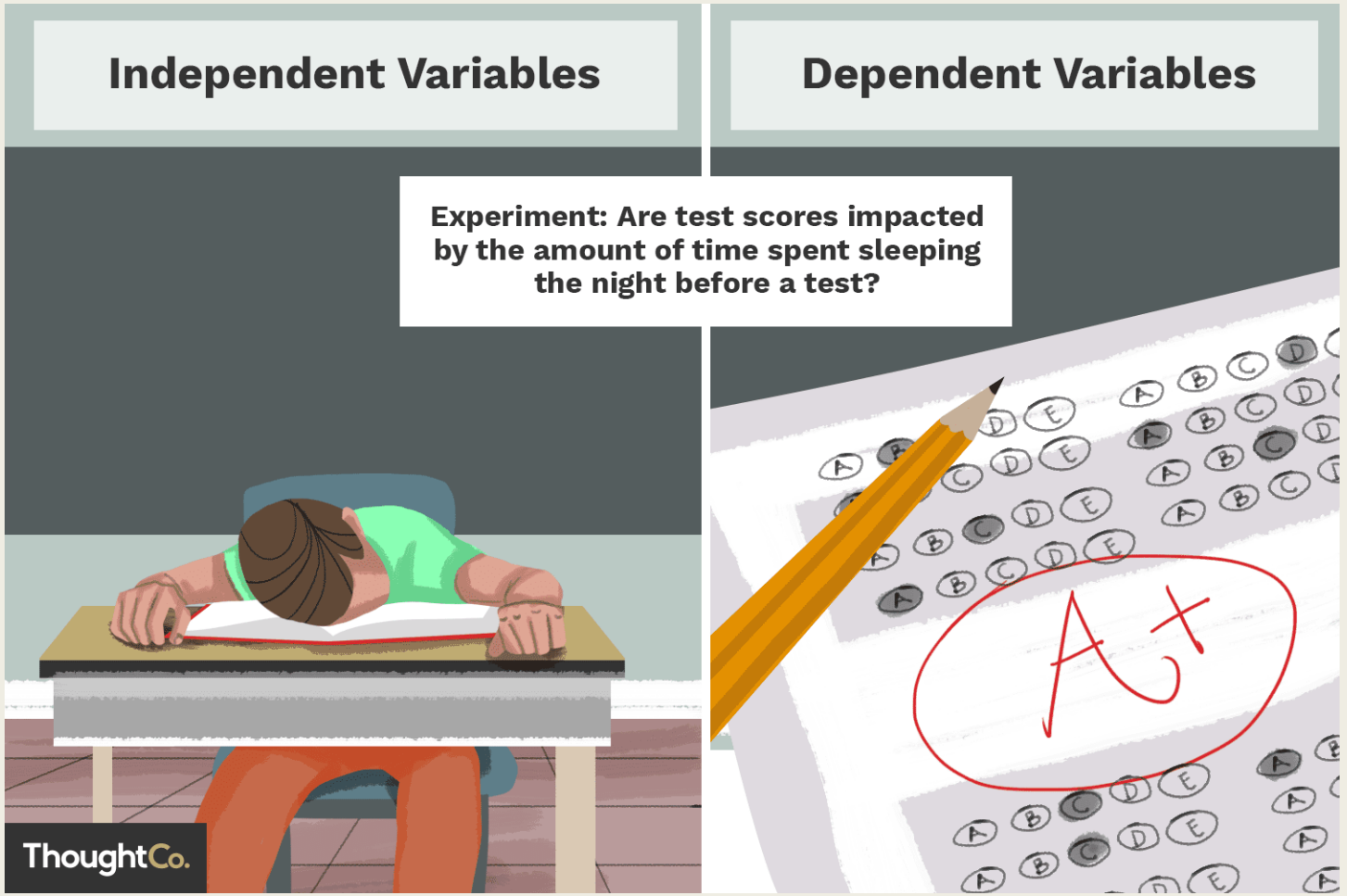
Variables

- A variable is a characteristic or properties of some event, object, or person that can take on different values or amounts.
- Most research begins with a general question about the relationship between two variables for a specific group of individuals.
- Variables may be…
  - Independent or dependent
  - Discrete or Continuous
  - Qualitative or Quantitative

Types of variables



Independent VS Dependent Variable



Independent and Dependent Variables

- When conducting research, experimenters often manipulate variables.

For example, an experimenter might compare the effectiveness of four types of antidepressants

- When a variable is manipulated by an experimenter, it is called an **independent variable**
- The experiment seeks to determine the effect of the independent variable on relief from depression.
- In this example, relief from depression is called a **dependent variable**.

Discrete and Continuous variables

- **Discrete variables** can take only certain values.

For example, a household could have three children or six children, but not 4.53 children

- **Continuous variables** can take any value within the range of the scale

For example, "time to respond to a question" are continuous variables since the scale is continuous and not made up of discrete steps, say, the response time could be 1.64 seconds

Qualitative and Quantitative variables

- **Qualitative variables** are those that express a qualitative attribute such as hair colour, eye colour, religion, favourite movie, and so on
  - The values of a qualitative variable do not imply a numerical ordering
  - For example: values of the variable "religion" differ qualitatively; no ordering of religions is implied.
  - Qualitative variables are also referred to as **categorical variables**
- **Quantitative variables** are those variables that are measured in terms of numbers.
  - Some examples of quantitative variables are height, weight, and shoe size.

Questions to clarify understanding: variables

Which of the following are qualitative variables?

1. Height measured in number of feet
2. Weight measured in number of kilograms
3. Number of days it snowed
4. Hair colour
5. Gender
6. Average daily temperature

Questions to clarify understanding: variables

In an experiment on the effect of sleep on memory, the **dependent variable** is

1 number of hours of sleep

2 recall score on a memory test

3 gender of the subjects

4 gender of the experimenter

Scale of measurement

Four fundamental scales

- Nominal
- Ordinal
- Interval
- Ratio

A summary of statistic for each scale measurement

Measure	Nominal	Ordinal	Interval and Ratio
Frequencies	Absolute Relative	Absolute Relative	Absolute Relative
Central Tendency	Mode	Mode, Median	# if you wish to treat outliers differently: Trimmed Mean # if distribution is symmetrical and outlier don't require different treatment: Mean # if distribution is skewed and outliers don't require different treatment Median, Model
Dispersion	Index Qualitative Variation (IQV)	Min/Max/Range/IQR	Min/Max/Range IQR Standard Deviation/Variance Coefficient of Variation
Graphs	Bar/Pie	Bar/Pie	Dot Plot Stem and leaf Histogram Box and Whisker Plot → determine skewness and kurtosis

The level of measurement and suggested inferential analysis

Independent Variable level	Dependent Variable level	Analysis
Dichotomous	Continuous	Independent Samples t-Test Linear Regression
Nominal or Ordinal	Continuous	ANOVA
Continuous	Continuous	Linear Regression Pearson's Correlation
Continuous or Categorical	Dichotomous	Binary Logistic Regression
Continuous or Categorical	Ordinal	Ordinal Logistic Regression
Categorical	Categorical	Chi-Square

Nominal Scales

- Nominal
  - Names or Categories
  - Examples include:
    - Gender
    - Handedness
    - Favourite colour
    - Religion
  - Lowest level of measurement

Ordinal Scales

- Ordinal:
  - Names or categories and order is meaningful
  - Example include
    - Consumer satisfaction rating
    - Military rank
    - Class ranking

Interval Scales

- Interval
  - Names or Categories, the order is meaningful, and intervals have the same interpretation
  - Example:
    - Celsius temperature scale
  - Problem: No true zero point

Ratio Scales

- Ratio:
  - Highest and most informative scale
  - Contains the qualities of nominal, ordinal, and interval scale **with the addition of an absolute zero point**
  - Example: amount of money-zero money indicates the absence of money

Question to clarify understanding: Measurement scales

Identify the scale of measurement for the following categorization of clothing: hat, shirt, shoes, pants

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Question to clarify understanding: Measurement scales

City of birth is an example of a(n)

1. Nominal scale
2. Ordinal scale
3. Interval scale
4. Ratio scale

Question to clarify understanding: Measurement scales

Identify the scale of measurement for the following: military title Lieutenant, Captain, Major.

1. nominal
2. ordinal
3. interval
4. ratio

Class Exercise

- Talk to your neighbour: Which level of measurement best describes the following measures?:
  - Telephone numbers
  - Gender
  - Participants' scores on an self-report anxiety question:
    - Strongly Agree (5), Agree(4), Neither Agree nor Disagree (3) Disagree(2), Strongly Disagree(1)
  - Height
  - University Rankings

Types of Descriptive Statistics

- Measures of Frequency:
- Measures of Central Tendency
- Measures of Dispersion or Variation
- Measures of Position
- Measures of Shape

Measures of Frequency

- Frequency
- Relative Frequency
- Cumulative Frequency

Example (1)

The advantage of relative frequency tables over frequency tables is that with percentages, you can compare categories.

	Chocolate	Strawberry	Vanilla	Total
Girls	50	40	60	150
Boys	65	30	70	165
Total	115	70	130	315

	Chocolate	Strawberry	Vanilla	Total
Girls	50/150=0.33	40/150=0.27	60/150=0.40	150/150=1.00
Boys	65/150=0.39	30/165=0.18	70/165=0.42	165/165=1.00
Total	115/315=0.37	70/315=0.22	130/315=0.41	315/315=1.00

Example (2)

you want to understand how frequently customers purchase your products that are priced up to 500 Baht.

Price Range (Baht)	Frequency	Relative Frequency	Cumulative Frequency
0-50	800	800/3,350=0.2388	800
51-100	1,200	1,200/3,350=0.3582	2,000
101-500	700	700/3,350=0.2090	2,700
501-1,000	450	450/3,350=0.1343	3,150
>1,000	200	200/3,350=0.0597	3,350
Total	3,350	1.0000	

Using the table above, you can easily identify that customers 2,700 times purchased products with prices up to 500 Baht.

Measures of Central Tendency

Measures of Central tendency: we use statistical measures to locate a single score that is most representative of all scores in a distribution.

- Mean
- Median
- Mode

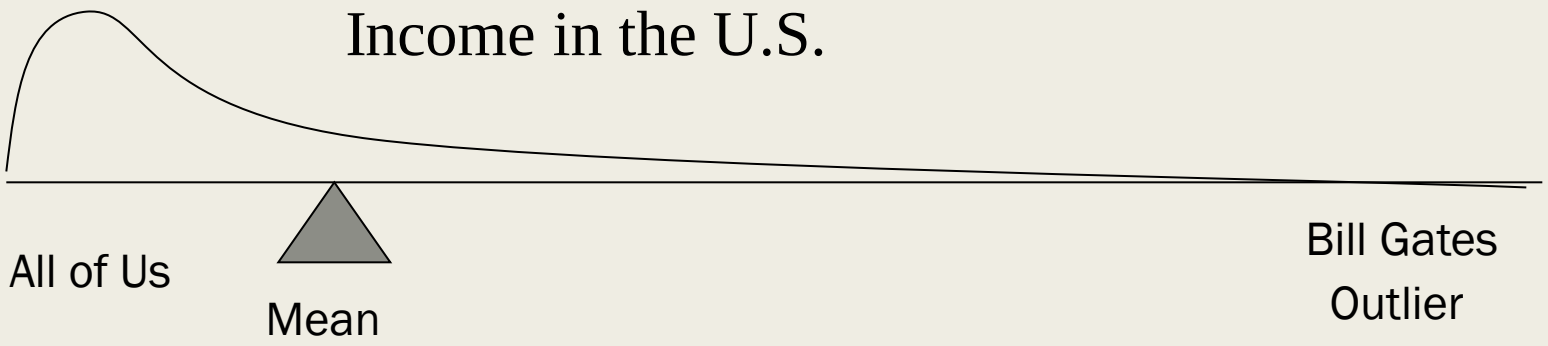
Mean

- To calculate the average of a set of observations, add their value and divide by the number of observations:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

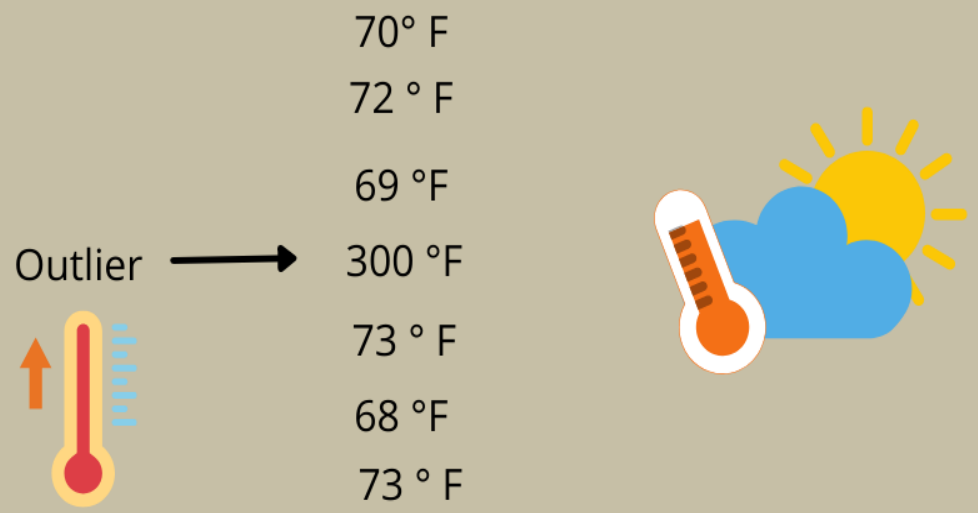
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Effects of outlier on Mean



Effects of outlier on Mean

Outliers influence the central tendency of the data.



What are Outliers?

Outliers are extreme behaviors. An outlier is a data point that differs significantly from other observations. It can cause serious problems in analysis.

Median

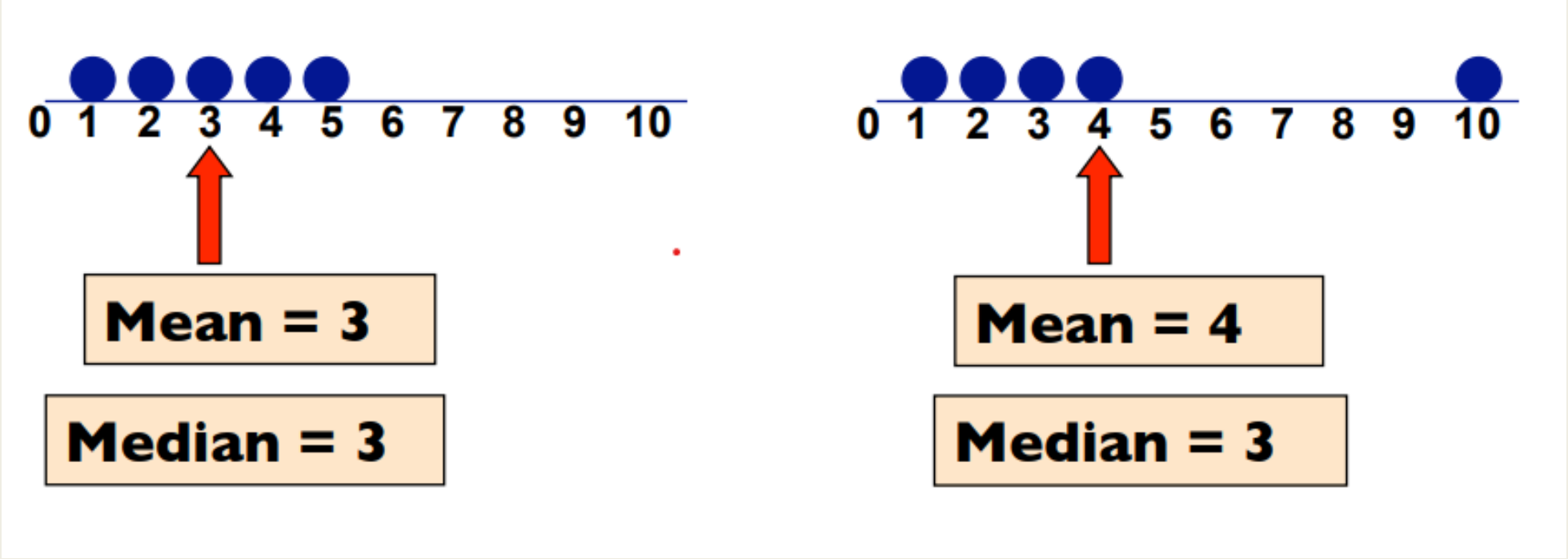
- Median – the exact middle value
- Calculation:
  - If there are an odd number of observations, find the middle value
  - If there are an even number of observations, find the middle two values and average them
- Example

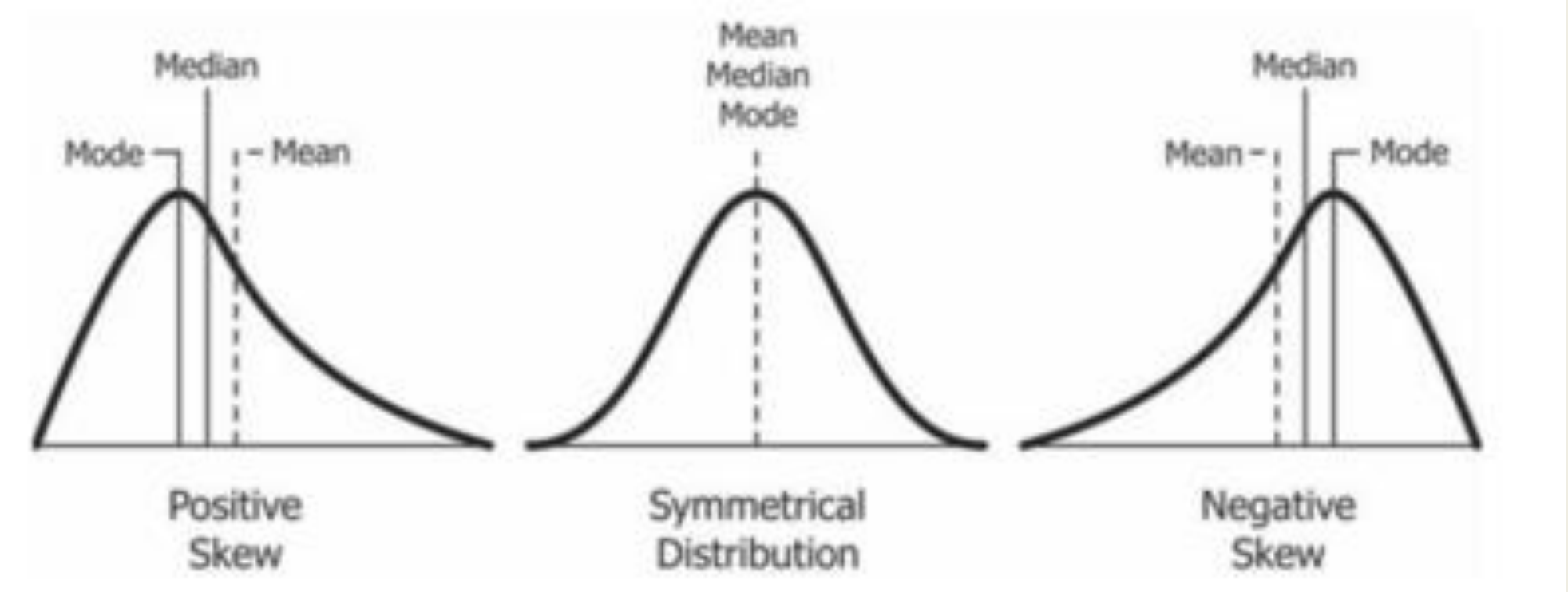
Data: 17 19 21 22 23 23 23 38

Median = (22+23)/2 = 22.5

Which Location Measure Is Best?

- Mean is best for symmetric distributions without outliers
- Median is useful for skewed distributions or data with outliers




Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Measures of Centrality: Type of Variable

	MODE	MEDIAN	MEAN
Nominal	Yes	No	No
Ordinal	Yes (Not Recommened)	Yes	No
Interval Ratio	Yes (Not Recommened)	Yes	Yes

Measures of Dispersion or Variation

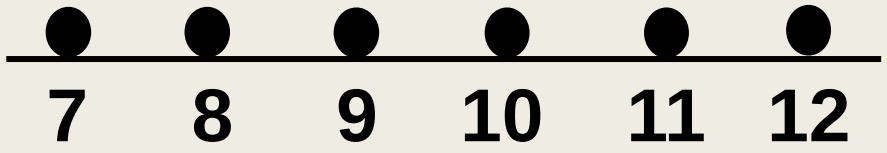
- Dispersion (Variability): a measure of the spread of scores in a distribution.
- Two sets of data have the same sample size, mean, and median, but they are different in terms of variability.
  - Range
  - Variance
  - Standard deviation

Range

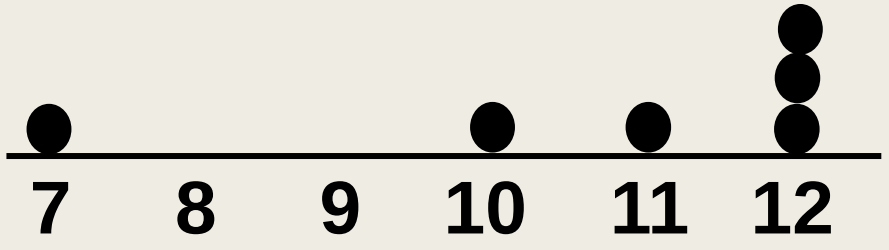
A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X_{max} and X_{min} are the two extreme observations then

$$\text{Range} = X_{max} - X_{min}$$

Range = 12 – 7 = 5

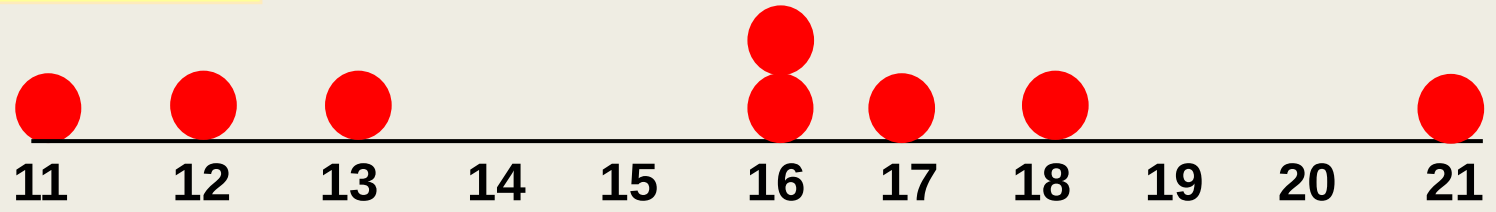


Range = 12 – 7 = 5



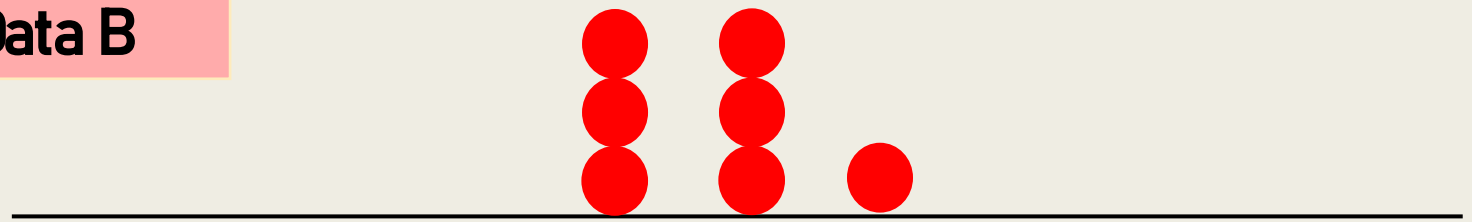
Standard Deviations

Data A



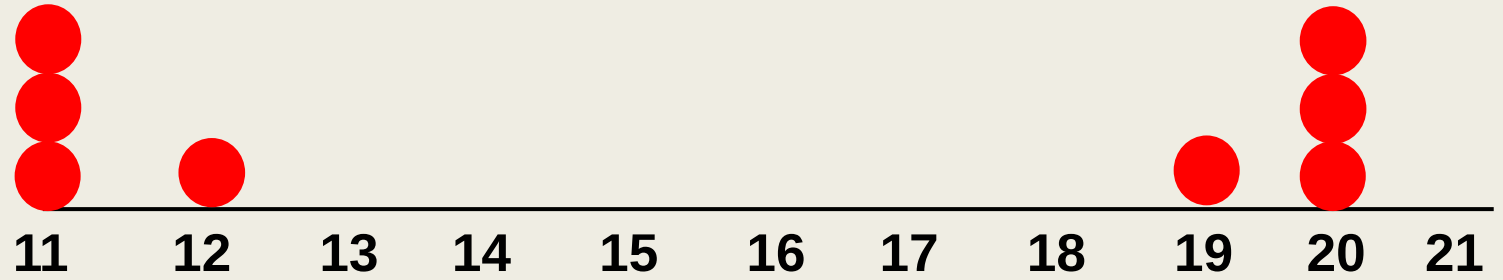
Mean = 15.5
s = 3.338

Data B

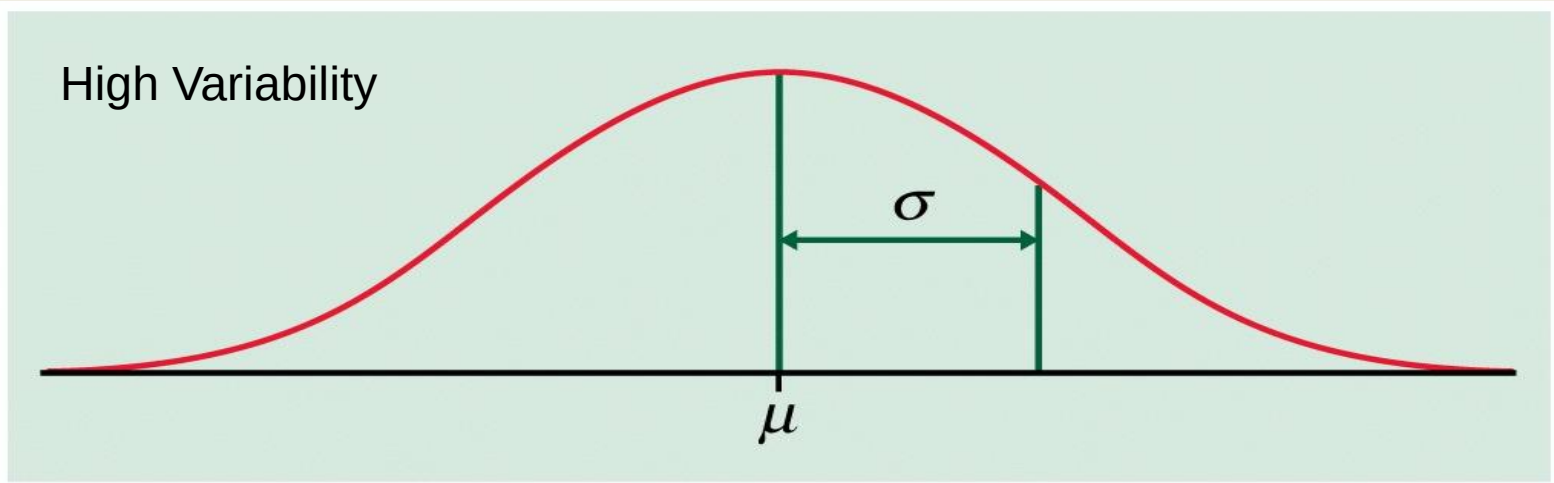


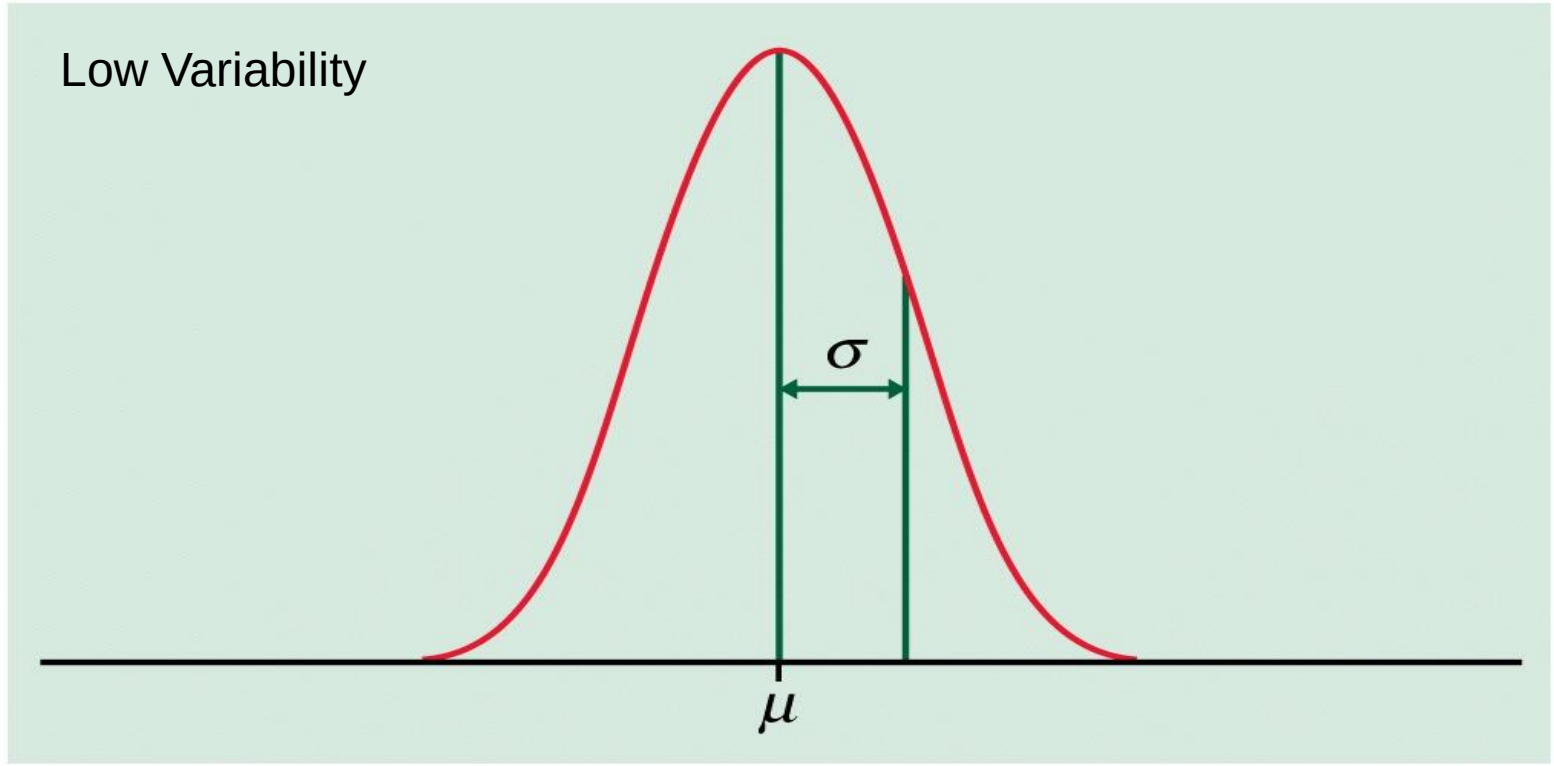
Mean = 15.5
s = .9258

Data C



Mean = 15.5
s = 4.57





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

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

Coefficient of Variation (CV)

- **The coefficient of variation (CV)** is a measure of relative variability. It is the ratio of the standard deviation to the mean (average). For example, the expression "The standard deviation is 15% of the mean" is a CV.
- The CV is particularly useful when you want to **compare results from two different surveys or tests** that have different measures or values. For example, if you are comparing the results from two tests that have different scoring mechanisms. If sample A has a CV of 12% and sample B has a CV of 25%, you would say that sample B has more variation, relative to its mean.

Coefficient of Variation (CV)

The formula for the coefficient of variation is:

Coefficient of Variation = (Standard Deviation/ Mean) * 100.

In symbols: $CV = \frac{S}{\bar{x}} \times 100$ or $CV = \frac{\sigma}{\mu} \times 100$

Multiplying the coefficient by 100 is an optional step to get a percentage, as opposed to a decimal.

Example

Problem Statement:

From the following data. Identify the risky project, is more risky:

Year	1	2	3	4	5
Project X (Cash profit in 1,000,000 Baht)	10	15	25	30	55
Project Y (Cash profit in 1,000,000 Baht)	5	20	40	40	30

Example

Solution:

In order to identify the risky project, we have to identify which of these projects is less consistent in yielding profits. Hence we work out **the coefficient of variation**.

Solution

Project X			Project y		
X	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	Y	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})^2$
10	-17	289	5	-22	484
15	-12	144	20	-7	49
25	-2	4	40	13	169
30	3	9	40	13	169
55	28	784	30	3	9
$\sum X = 135$		$\sum (X_i - \overline{X})^2 = 1230$	$\sum Y = 135$		$\sum (Y_i - \overline{Y})^2 = 880$

Project X

Here $\overline{X} = \frac{\sum X}{n}$

$= \frac{135}{5}$

$= 27$

$s = \sqrt{\frac{\sum (X - \overline{X})}{n-1}}$

$= \sqrt{\frac{1230}{4}}$

$= 17.54$

$CV = 0.65$

Project Y

Here $\overline{Y} = \frac{\sum Y}{n}$

$= \frac{135}{5}$

$= 27$

$s = \sqrt{\frac{\sum (X - \overline{X})}{n-1}}$

$= \sqrt{\frac{880}{4}}$

$= 14.83$

$CV = 0.55$

Since CV is higher for project X than for project Y, hence despite the average profits being same, **project X is more risky**.

IQV—Index of Qualitative Variation

- For nominal variables
- Statistic for determining the dispersion of cases across categories of a variable.
- Ranges from 0 (no dispersion or variety) to 1 (maximum dispersion or variety)
- 1 refers to even numbers of cases in all categories, NOT that cases are distributed like population proportions
- IQV is affected by the number of categories

To calculate:

$$IQV = \frac{K(100^2 - \sum category\%^2)}{100^2(K-1)}$$

K=# of categories

Cat.% = percentage in each category

Problem: Is SJSU more diverse than UC Berkeley?

Solution: Calculate IQV for each campus to determine which is higher.

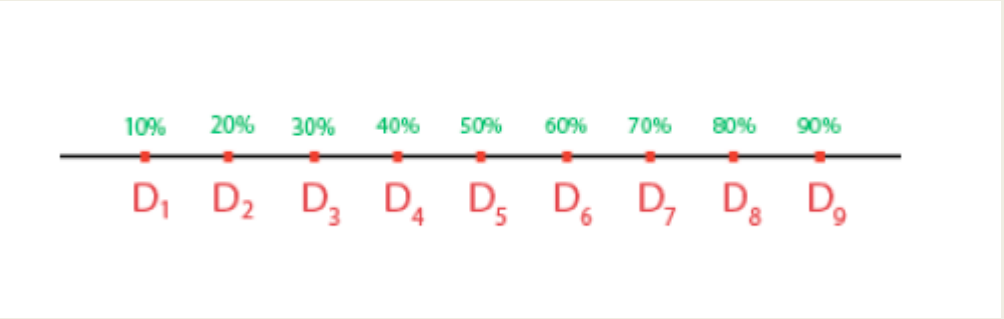
SJSU:		UC Berkeley:	
Percent	Category	Percent	Category
00.6	Native American	00.6	Native American
06.1	Black	03.9	Black
39.3	Asian/PI	47.0	Asian/PI
19.5	Latino	13.0	Latino
34.5	White	35.5	White

What can we say before calculating? Which campus is more evenly distributed?

SJSU IQV = 0.856 and **UCB IQV = 0.793**

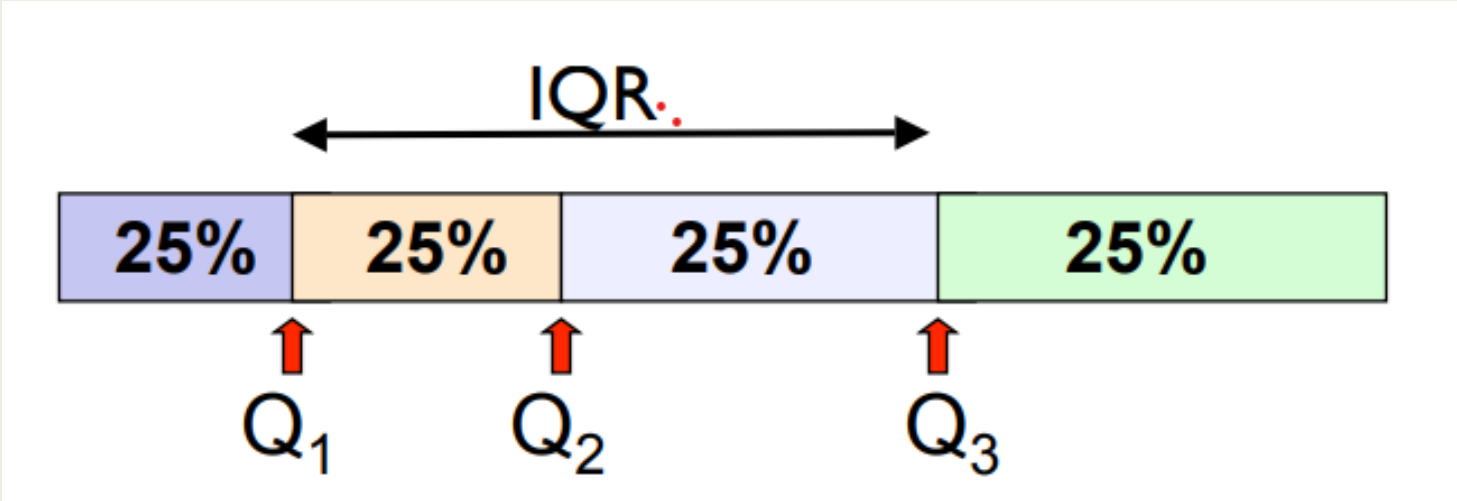
Measures of Position

- Quartile
- Decile



- Percentile

Quartiles and IQR



- Split the series in 4 equal parts.
- The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
- Q2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

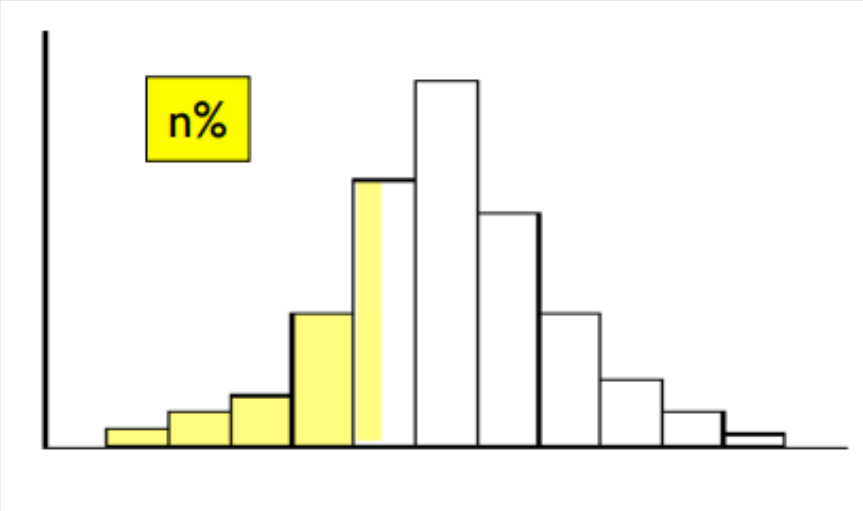
Why Percentiles?

- Raw data is sometimes difficult to interpret by itself.
- For example, your test score is 35/50. Is this good or bad result

- Percentile shows what proportion of score is higher that your score
- Suppose your score is higher that 65% of the population, that is your score is the 65^{th} percentiles.

Definition: Percentiles

- Split the series in 100 equal parts
- In general the nth percentile is a value such that n% of the observations fall at or below or it



First quartile (designated Q1) also called the lower quartile = the 25th percentile

Second quartile (designed Q2) also called the Median = the 50th percentile

Third quartile (designed Q3) also called the upper quartile = the 75th percentile

Computation: Percentiles

The i - th percentile of a list of N ordered values (sorted from least to greatest) is the score computed according the following method:

1. Compute the ordinal rank by the formula

$$R = \frac{i}{100} \times (N + 1)$$

If R is an integer then i is the score with the index R, otherwise take I_R and F_R as an integer part and fractional part of R respectively;

2. Let s_1 is the score with index I_R and s2 the score with index I_R +1;

3. Compute $i = s_1 + F_R(s_2 - s_1)$

Example: 25th percentile of 20 Quiz scores

Score	Rank
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

$$R = \frac{25}{100} \times (20+1)$$
$$= \frac{21}{4}$$
$$= 5.25$$

and therefore $l_R = 5$, $F_R = 0.25$, $s_1 = 5$ and $s_2 = 5$.

The 25th percentile is 5+0.25(5–5) = 5

Example: 85th percentile of 20 Quiz scores

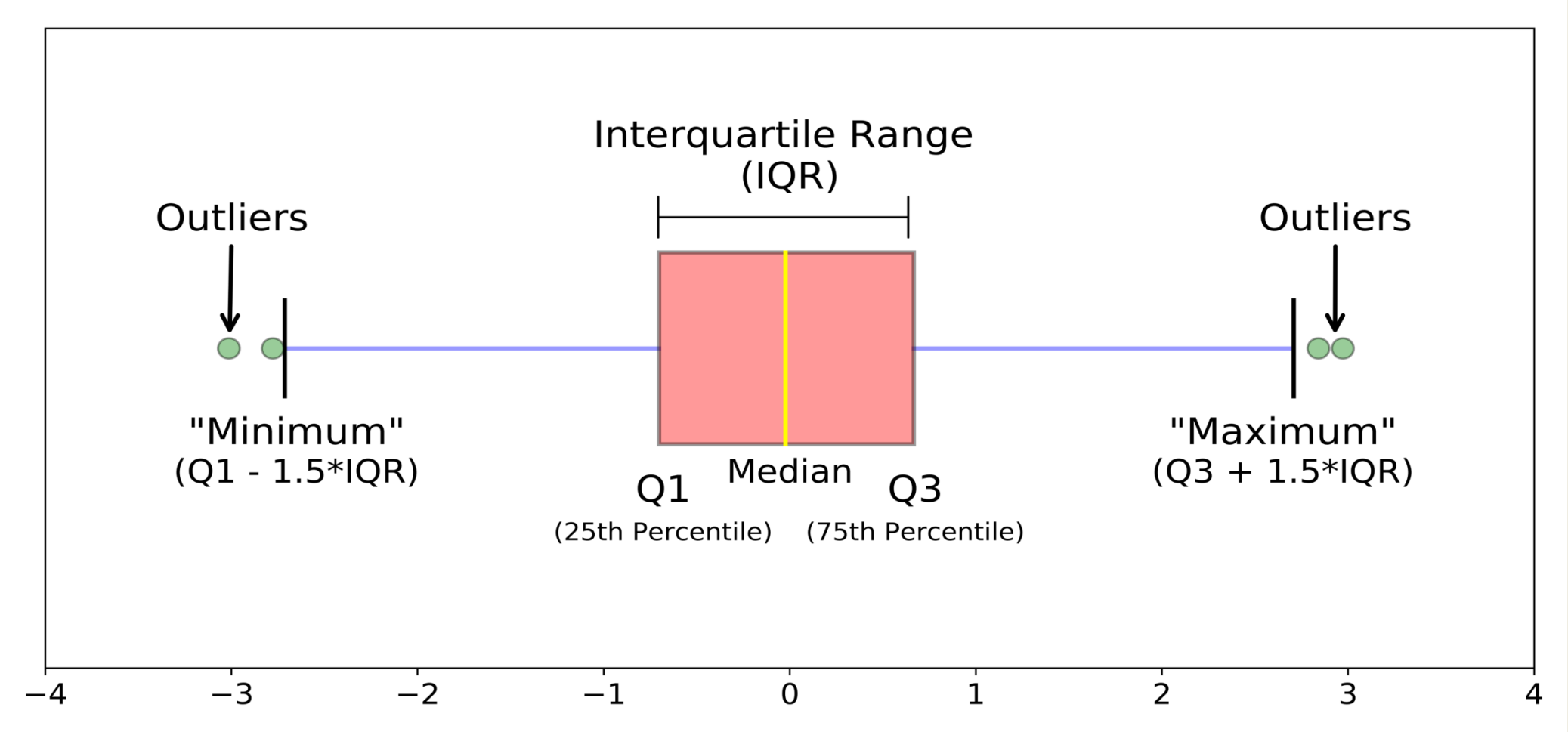
Score	Rank
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

$$R = \frac{85}{100} \times (20+1)$$
$$= 0.85 \times 21$$
$$= 17.85$$

and therefore $l_R = 17$, $F_R = 0.85$, $s_1 = 9$ and $s_2 = 10$.

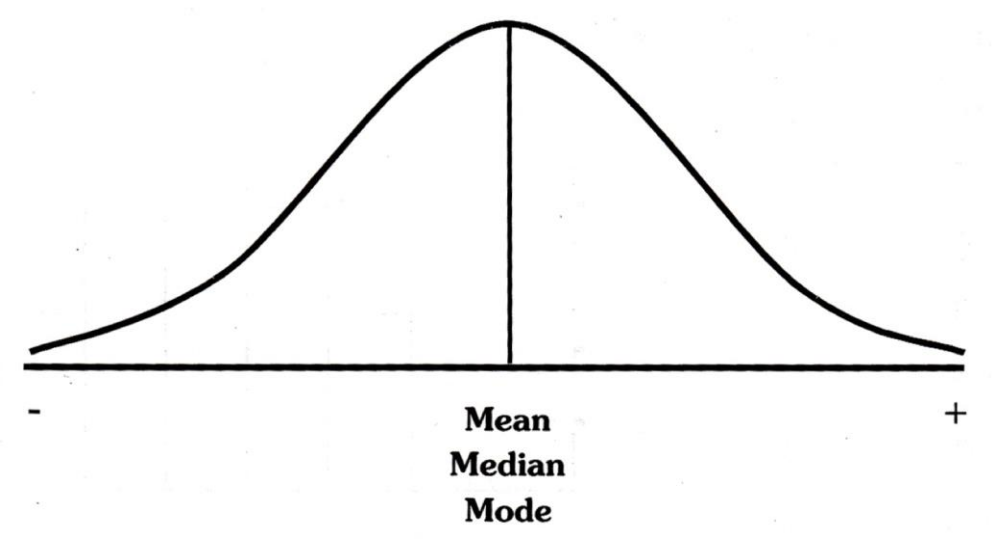
The 85th percentile is 9+0.85(10–9) = 9.85

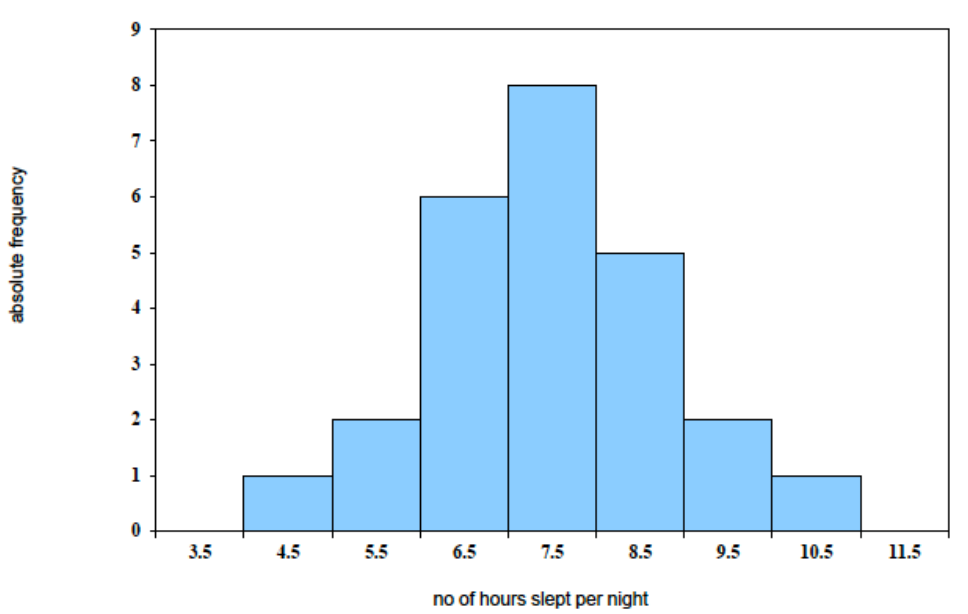
Boxplots


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Measures of Shape

- Symmetry





Measures of Shape

- Skewness
- Kurtosis

Skewness

- Indicate for a series of data:
  - *Deviation from the symmetry*
  - *Direction of the deviation from symmetry (positive/negative)*
- Formula for calculus:

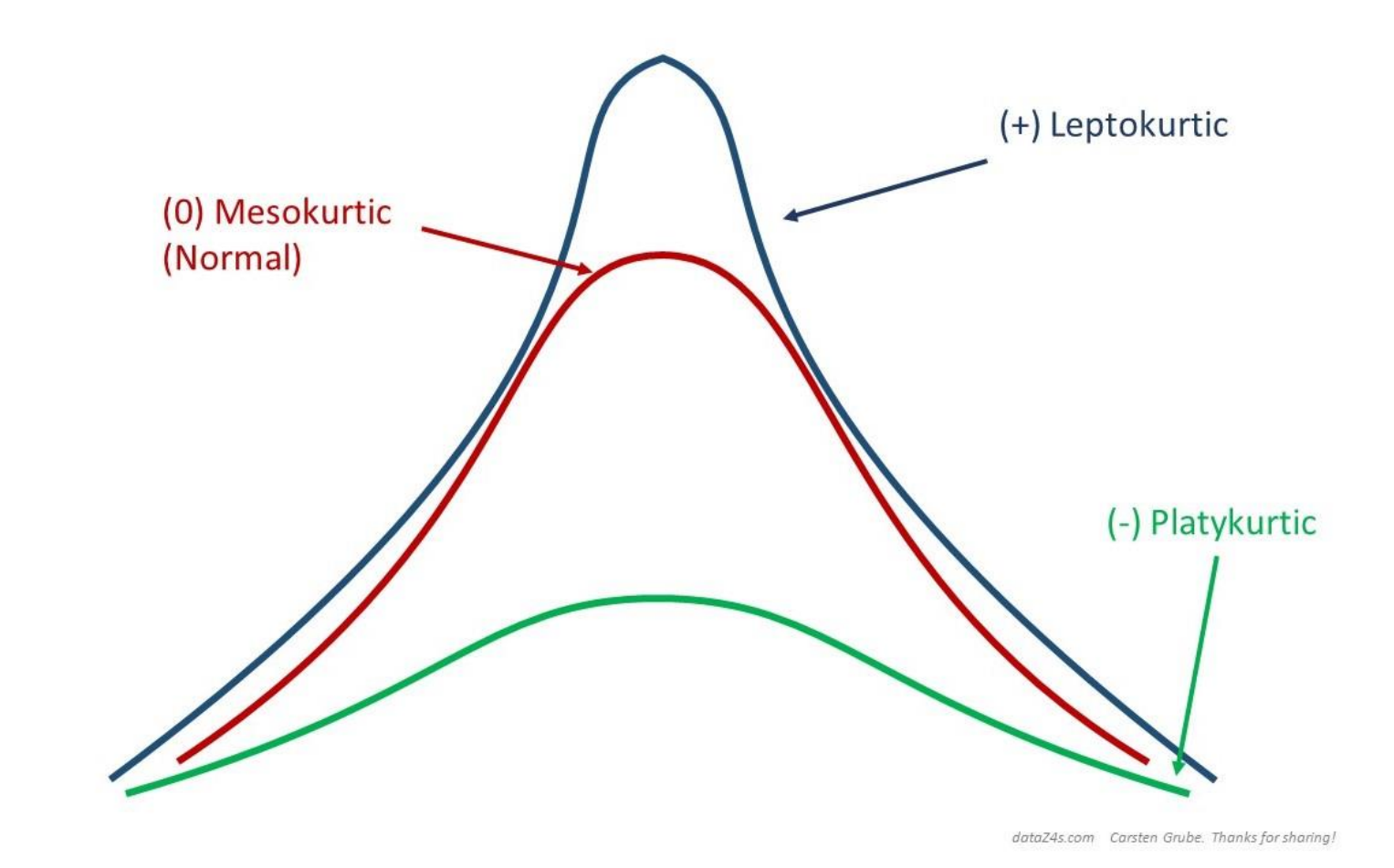
$$M_3 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^3}{n}$$

Kurtosis

- a measurement about the extremities (i.e. tails) of the distribution of data, and therefore provides an indication of the presence of outliers.
- Formula for calculus:

$$\alpha_4 = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^4}{S^4} - 3$$

The degrees of kurtosis

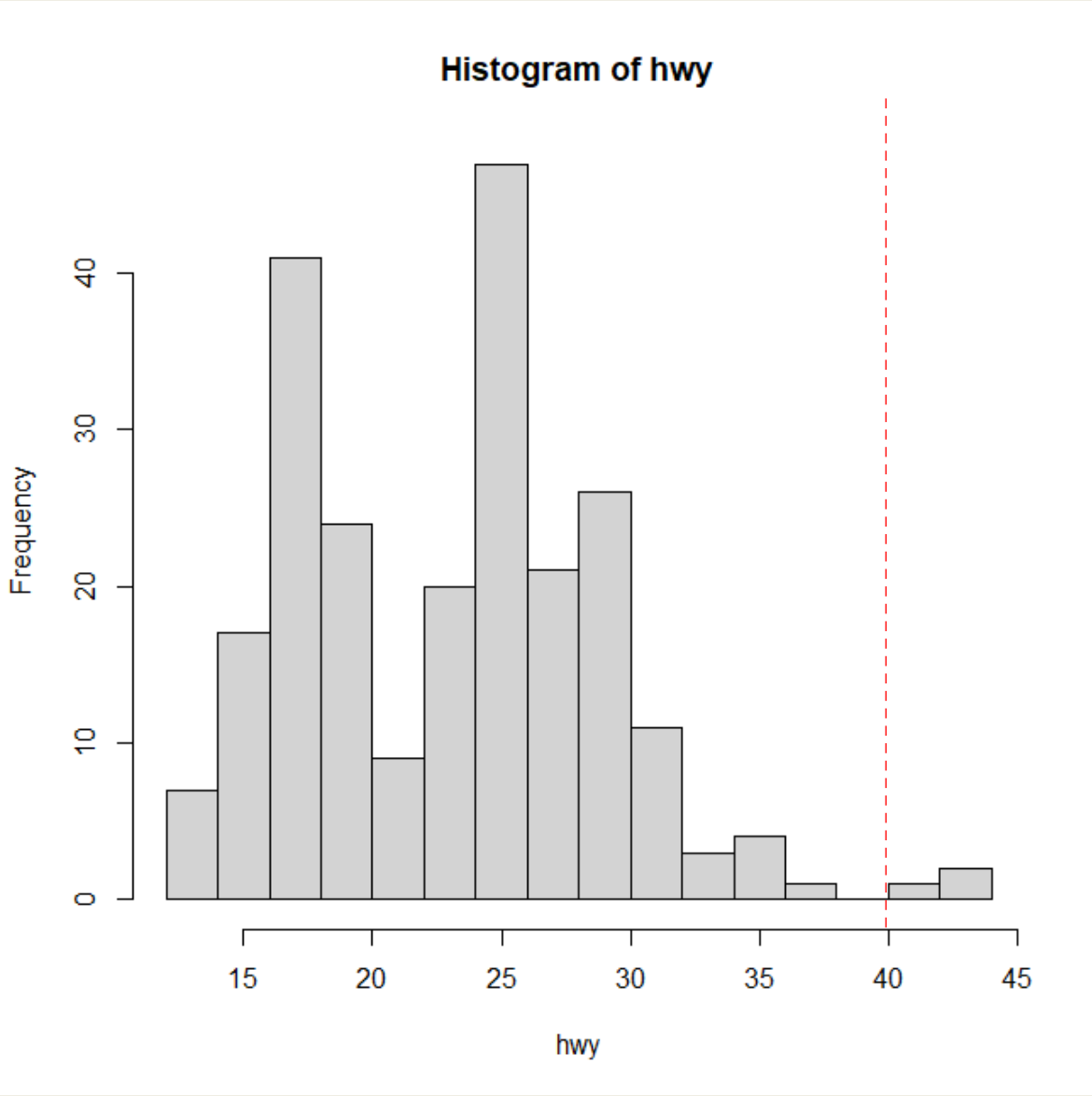


How to detect outlier

- Histogram
- Boxplot
- Percentiles
- Hampel filter
- Statistical tests: Grubbs's test, Dixon's test, Rosner's test

Histogram

- A basic way to detect outliers is to draw a histogram of the data.
- Using R base (with the number of bins corresponding to the square root of the number of observations in order to have more bins than the default option):

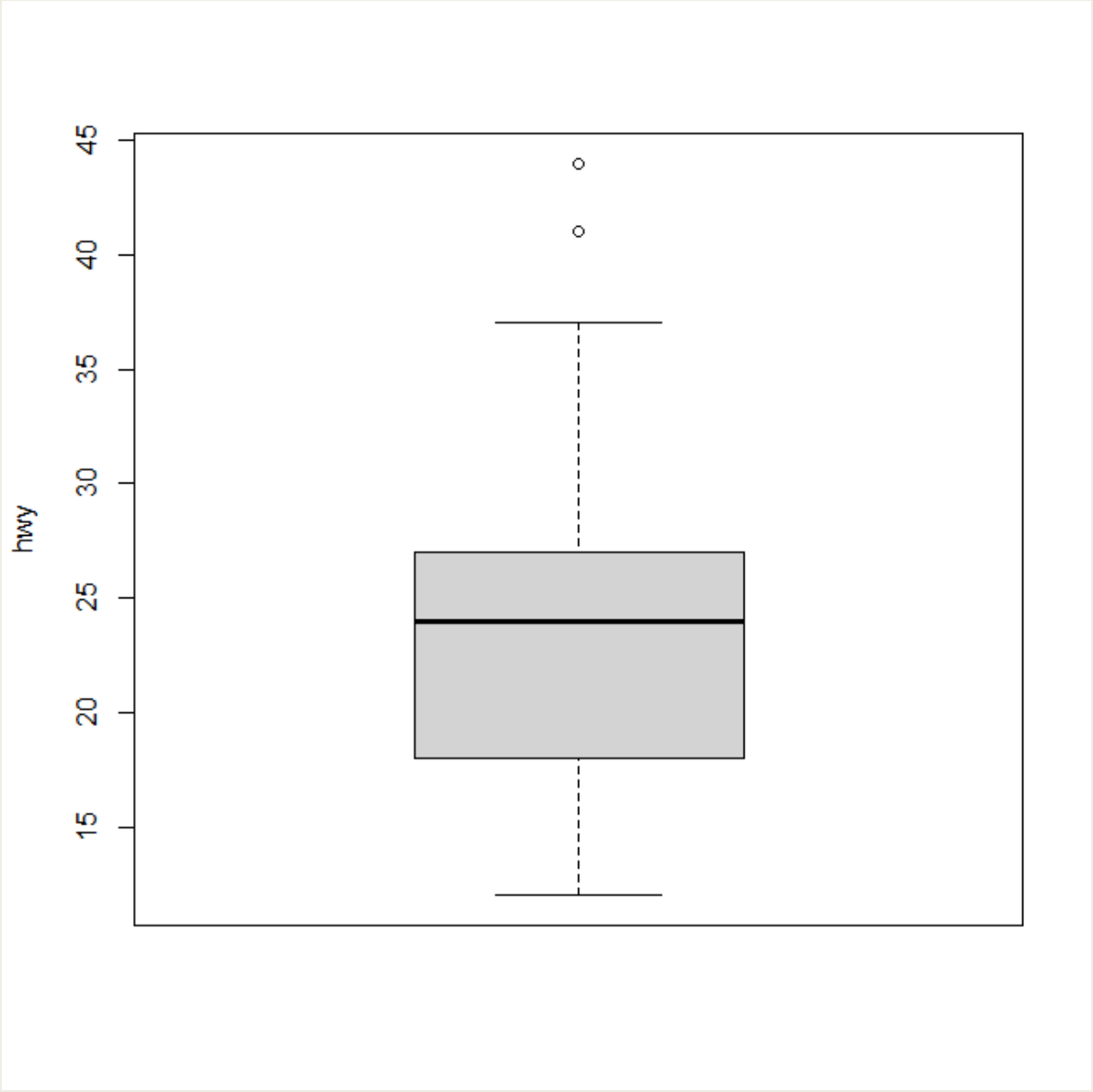


Boxplot

All observations outside of the following interval will be considered as potential outliers:

$$I=[Q_1-1.5\cdot IQR;Q_3+1.5\cdot IQR]$$

It is also possible to extract the values of the potential outliers based on the IQR criterion thanks to the boxplot.stats()$out function:



Percentiles

With the percentiles method, all observations that lie outside the interval formed by the 2.5 and 97.5 percentiles will be considered as potential outliers.

The values of the lower and upper percentiles (and thus the lower and upper limits of the interval) can be computed with the quantile() function.

Hampel filter

- Outliers considered as the values outside the interval (I) formed by the median, plus or minus 3 median absolute deviations (MAD).
- The values of the lower and upper percentiles (and thus the lower and upper limits of the interval) can be computed with the quantile() function.

$$I=[\text{median}-3\cdot MAD;\text{median}+3\cdot MAD]$$

where MAD is the median absolute deviation and is defined as the median of the absolute deviations from the data's median ($\overline{X}=\text{median}(X)$):

$$MAD=\text{median}(|X_i-\overline{X}|)$$

Statistical tests: Grubbs's test

- The Grubbs test detects one outlier at a time (highest or lowest value), so the null and alternative hypotheses are as follows:

  H_0: The *highest* value is **not** an outlier

  H_1: The *highest* value is an outlier

if we want to test the highest value, or:

  H_0: The *lowest* value is **not** an outlier

  H_1: The *lowest* value is an outlier

if we want to test the lowest value.

- Note that the Grubbs test is not appropriate for sample size of 6 or less ($n \leq 6$).

Statistical tests: Grubbs's test

- If you suspect that the maximum value in the data set may be an outlier you can use the following test statistic.

$$G = \frac{x_{max} - \bar{x}}{s}$$

- If you suspect that the minimum value in the data set may be an outlier you can use the following test statistic

$$G = \frac{\bar{x} - x_{min}}{s}$$

- If the ***p*-value** **is less** than the chosen **significance threshold** (generally α=0.05) then the null hypothesis is rejected, we will conclude that the **lowest/highest value is an outlier**.
- On the contrary, if the ***p*-value** **is greater or equal** than the **significance level**, the null hypothesis is not rejected, and we will conclude that, based on the data, we do not reject the hypothesis that the **lowest/highest value is not an outlier**.

Statistical tests: Dixon's test

- Similar to the Grubbs test, Dixon test is used to test whether a single low or high value is an outlier. So if more than one outliers is suspected, the test has to be performed on these suspected outliers individually.
- To perform the Dixon's test in R, we use the dixon.test() function from the {outliers}
- Note that Dixon test is most useful for small sample size (usually $n \leq 25$).

Statistical tests: Rosner's test

- Rosner's test for outliers has the advantages that:
  - *it is used to detect several outliers at once (unlike Grubbs and Dixon test which must be performed iteratively to screen for multiple outliers), and*
  - *it is designed to avoid the problem of masking, where an outlier that is close in value to another outlier can go undetected.*
- Unlike Dixon test, note that Rosner test is most appropriate when the sample size is large ($n \geq 20$).
- To perform the Rosner test we use the rosnerTest() function from the {EnvStats} package.

References

- https://statsandr.com/blog/outliers-detection-in-r/
- https://www.statology.org/grubbs-test-r/