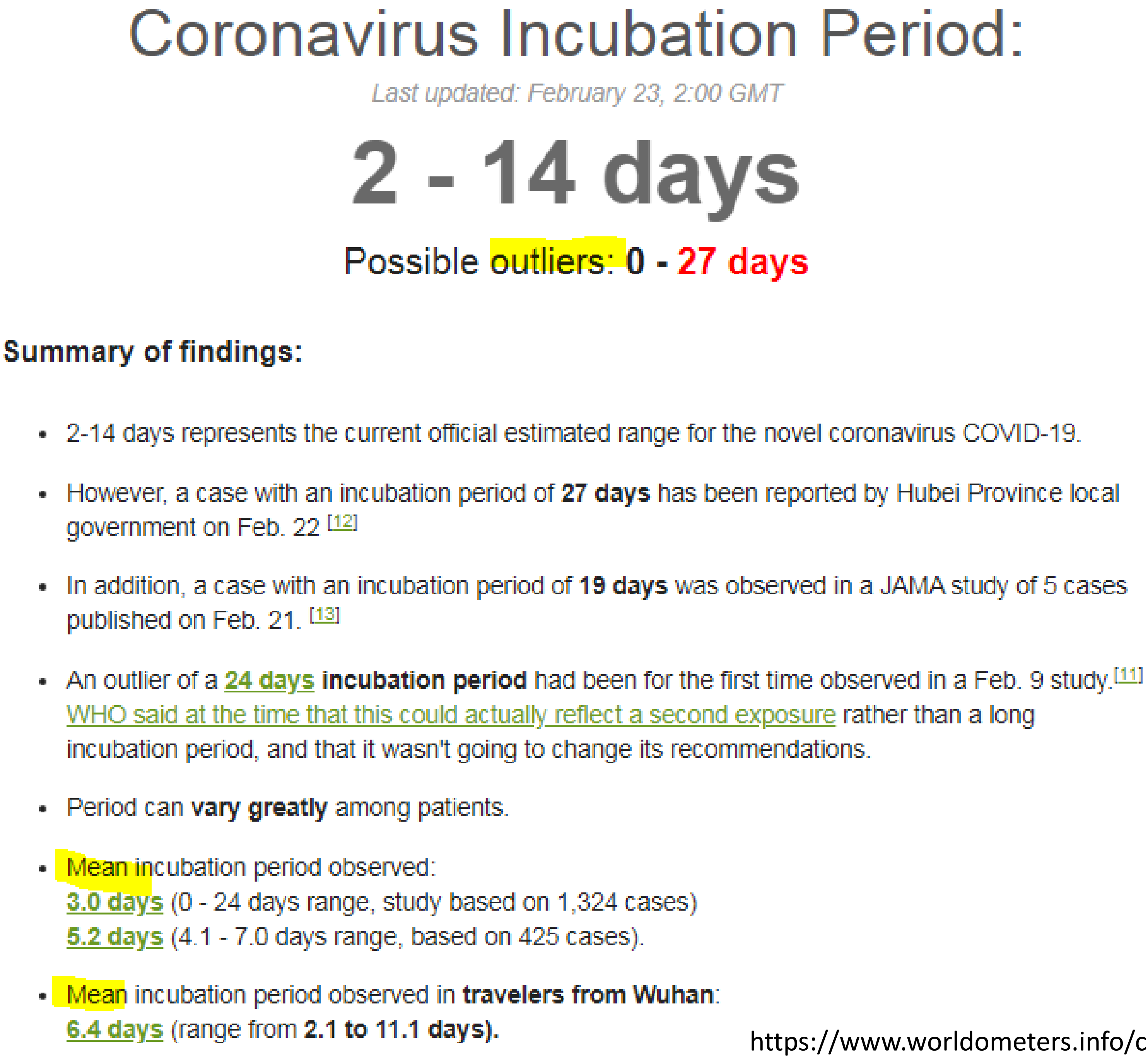
Descriptive statistics

Learning Objective (LO)

LO1: to identify parameters in descriptive statistics

LO2: to summarize/ describe data

LO3: To illustrate/ report data



Coronavirus Incubation Period:

Last updated: February 23, 2:00 GMT

2 - 14 days

Possible outliers: 0 - **27 days**

Summary of findings:

- 2-14 days represents the current official estimated range for the novel coronavirus COVID-19.
- However, a case with an incubation period of **27 days** has been reported by Hubei Province local government on Feb. 22 [12]
- In addition, a case with an incubation period of **19 days** was observed in a JAMA study of 5 cases published on Feb. 21. [13]
- An outlier of a **24 days incubation period** had been for the first time observed in a Feb. 9 study.[11] WHO said at the time that this could actually reflect a second exposure rather than a long incubation period, and that it wasn't going to change its recommendations.
- Period can **vary greatly** among patients.
- Mean incubation period observed:
  3.0 days (0 - 24 days range, study based on 1,324 cases)
  5.2 days (4.1 - 7.0 days range, based on 425 cases).
- Mean incubation period observed in **travelers from Wuhan**:
  6.4 days (range from **2.1 to 11.1 days).**

https://www.worldometers.info/coronavirus/

Days from first symptom to death

The Wang et al. February 7 study published on JAMA found that the median time from first symptom to dyspnea was 5.0 days, to hospital admission was 7.0 days, and to ARDS was 8.0 days.[9]

Previously. the China National Health Commission reported the details of the first 17 deaths up to 24 pm 22 Jan 2020. A study of these cases found that the median **days from first symptom to death were 14** (range 6-41) days, and tended to be shorter among people of 70 year old or above (11.5 [range 6-19] days) than those with ages below 70 year old (20 [range 10-41] days.[6]

Median Hospital Stay

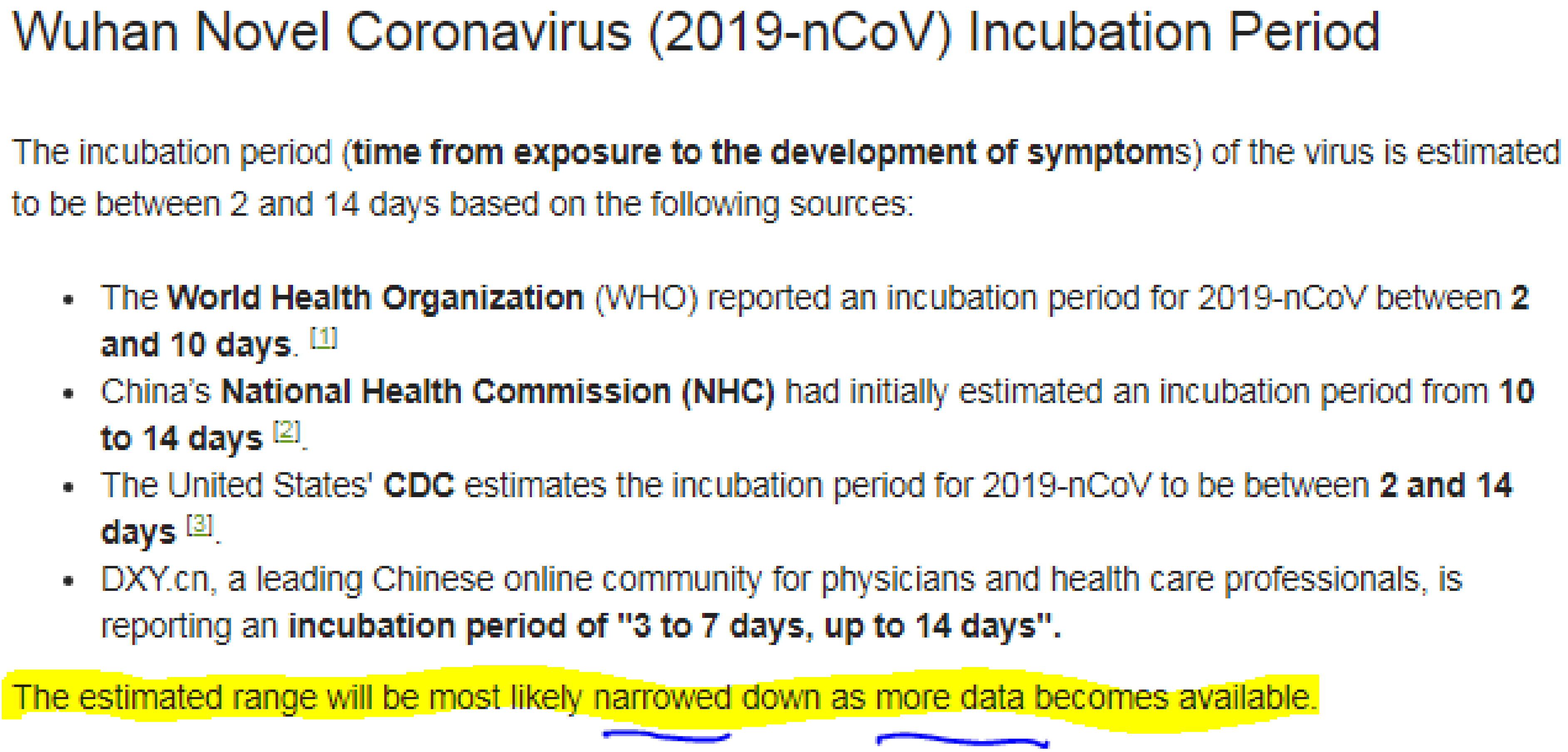
The JANA study found that, among those discharged alive, the **median hospital stay was 10 days**.[9]

Comparison with other viruses

For comparison, the case fatality rate with seasonal flu in the United States is less than 0.1% (1 death per every 1,000 cases).

Mortality rate for SARS was 10%, and for MERS 34%.

https://www.worldometers.info/coronav



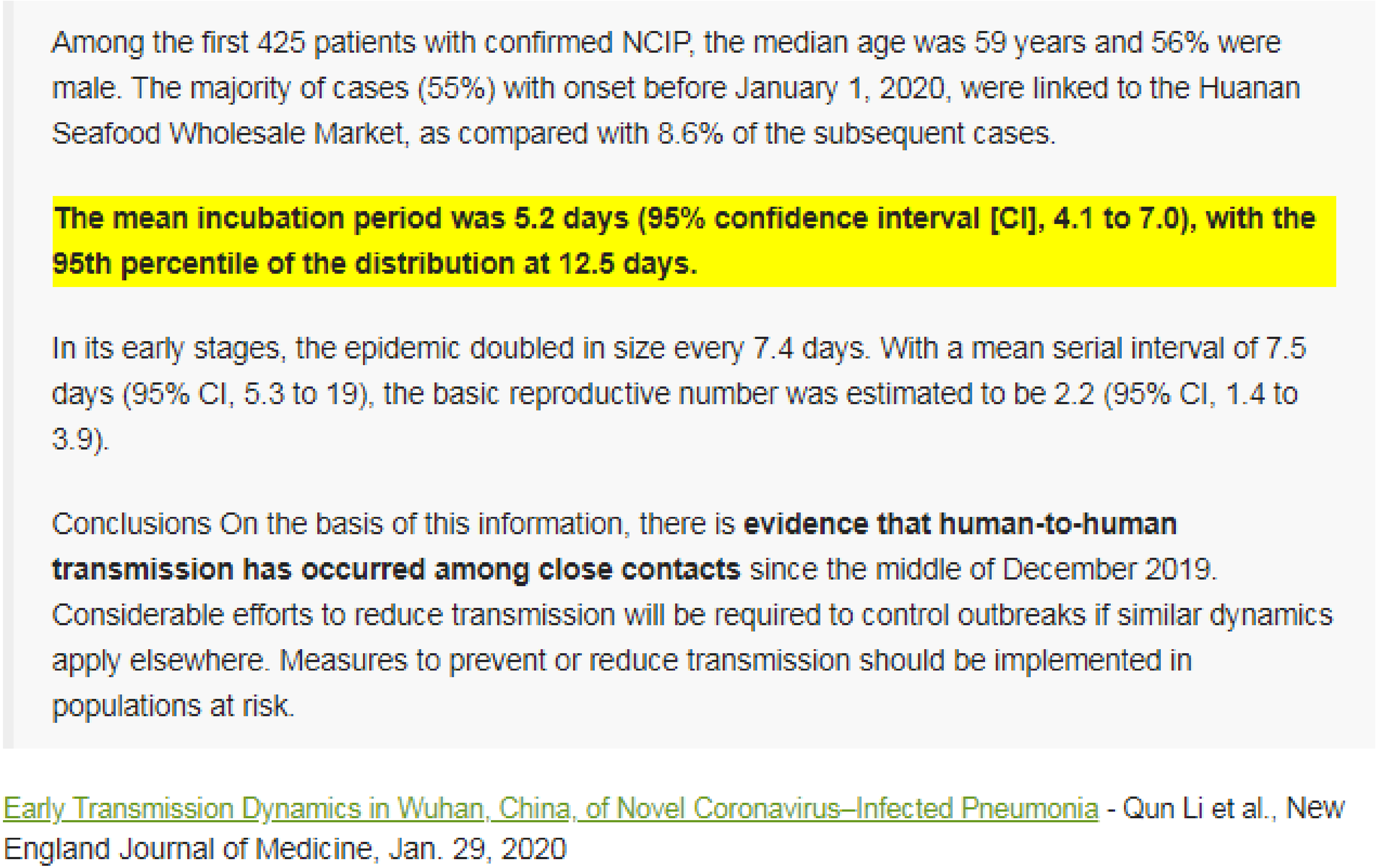
Wuhan Novel Coronavirus (2019-nCoV) Incubation Period

The incubation period (**time from exposure to the development of symptoms**) of the virus is estimated to be between 2 and 14 days based on the following sources:

- The **World Health Organization** (WHO) reported an incubation period for 2019-nCoV between **2 and 10 days**. [1]
- China's **National Health Commission (NHC)** had initially estimated an incubation period from **10 to 14 days** [2].
- The United States' **CDC** estimates the incubation period for 2019-nCoV to be between **2 and 14 days** [3].
- DXY.cn, a leading Chinese online community for physicians and health care professionals, is reporting an **incubation period of "3 to 7 days, up to 14 days".**

The estimated range will be most likely narrowed down as more data becomes available.

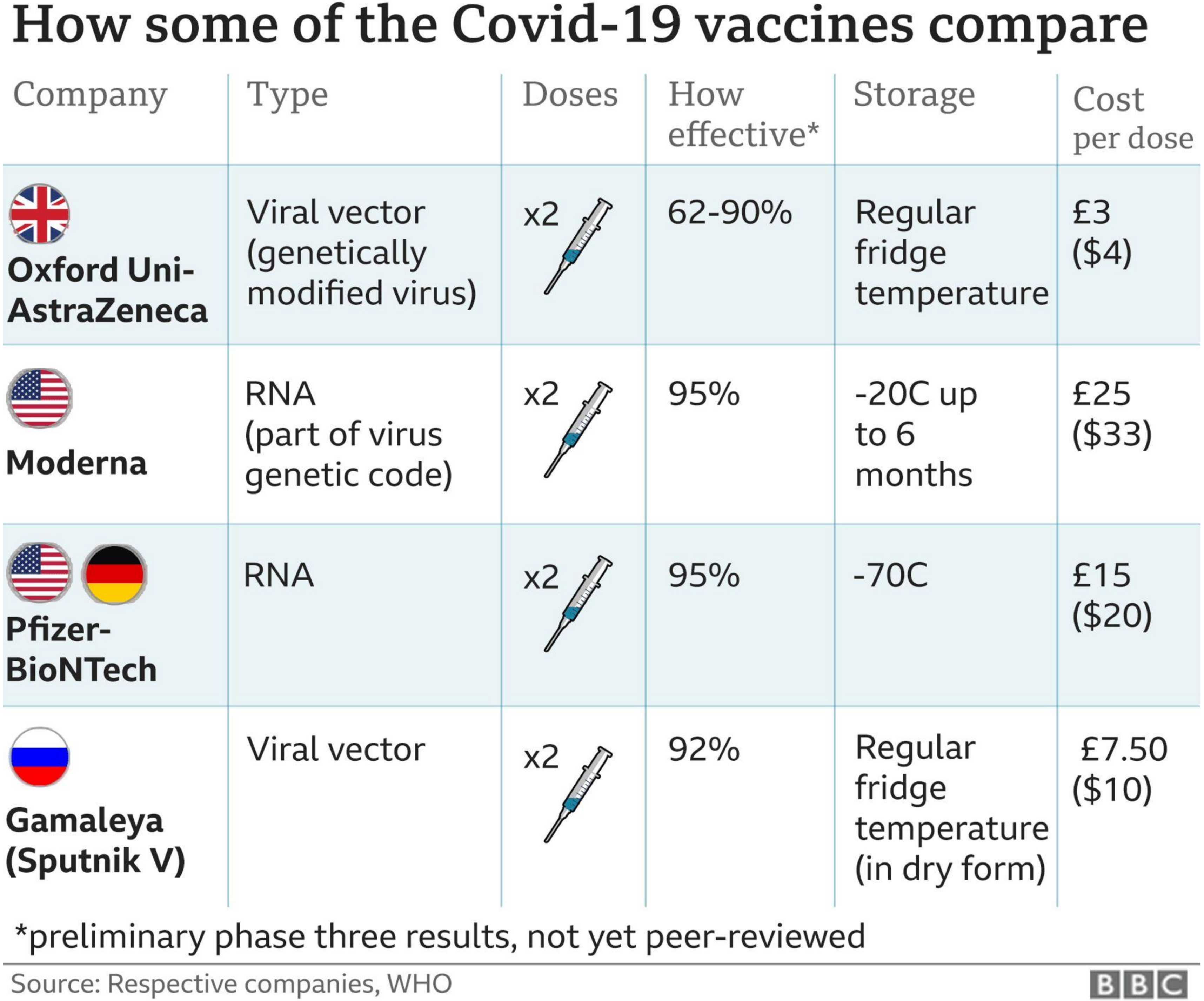
https://www.worldometers.info/coronavirus/



> Among the first 425 patients with confirmed NCIP, the median age was 59 years and 56% were male. The majority of cases (55%) with onset before January 1, 2020, were linked to the Huanan Seafood Wholesale Market, as compared with 8.6% of the subsequent cases.
>
> **The mean incubation period was 5.2 days (95% confidence interval [CI], 4.1 to 7.0), with the 95th percentile of the distribution at 12.5 days.**
>
> In its early stages, the epidemic doubled in size every 7.4 days. With a mean serial interval of 7.5 days (95% CI, 5.3 to 19), the basic reproductive number was estimated to be 2.2 (95% CI, 1.4 to 3.9).
>
> Conclusions On the basis of this information, there is **evidence that human-to-human transmission has occurred among close contacts** since the middle of December 2019. Considerable efforts to reduce transmission will be required to control outbreaks if similar dynamics apply elsewhere. Measures to prevent or reduce transmission should be implemented in populations at risk.

Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia - Qun Li et al., New England Journal of Medicine, Jan. 29, 2020

https://www.worldometers.info/coronavirus/



How some of the Covid-19 vaccines compare

Company	Type	Doses	How effective*	Storage	Cost per dose
Oxford Uni-AstraZeneca	Viral vector (genetically modified virus)	x2	62-90%	Regular fridge temperature	£3 ($4)
Moderna	RNA (part of virus genetic code)	x2	95%	-20C up to 6 months	£25 ($33)
Pfizer-BioNTech	RNA	x2	95%	-70C	£15 ($20)
Gamaleya (Sputnik V)	Viral vector	x2	92%	Regular fridge temperature (in dry form)	£7.50 ($10)

*preliminary phase three results, not yet peer-reviewed

Source: Respective companies, WHO

BBC

Statistical Methods

- Descriptive statistics
  - Collecting and describing data
- Inferential statistics
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

Descriptive Statistics

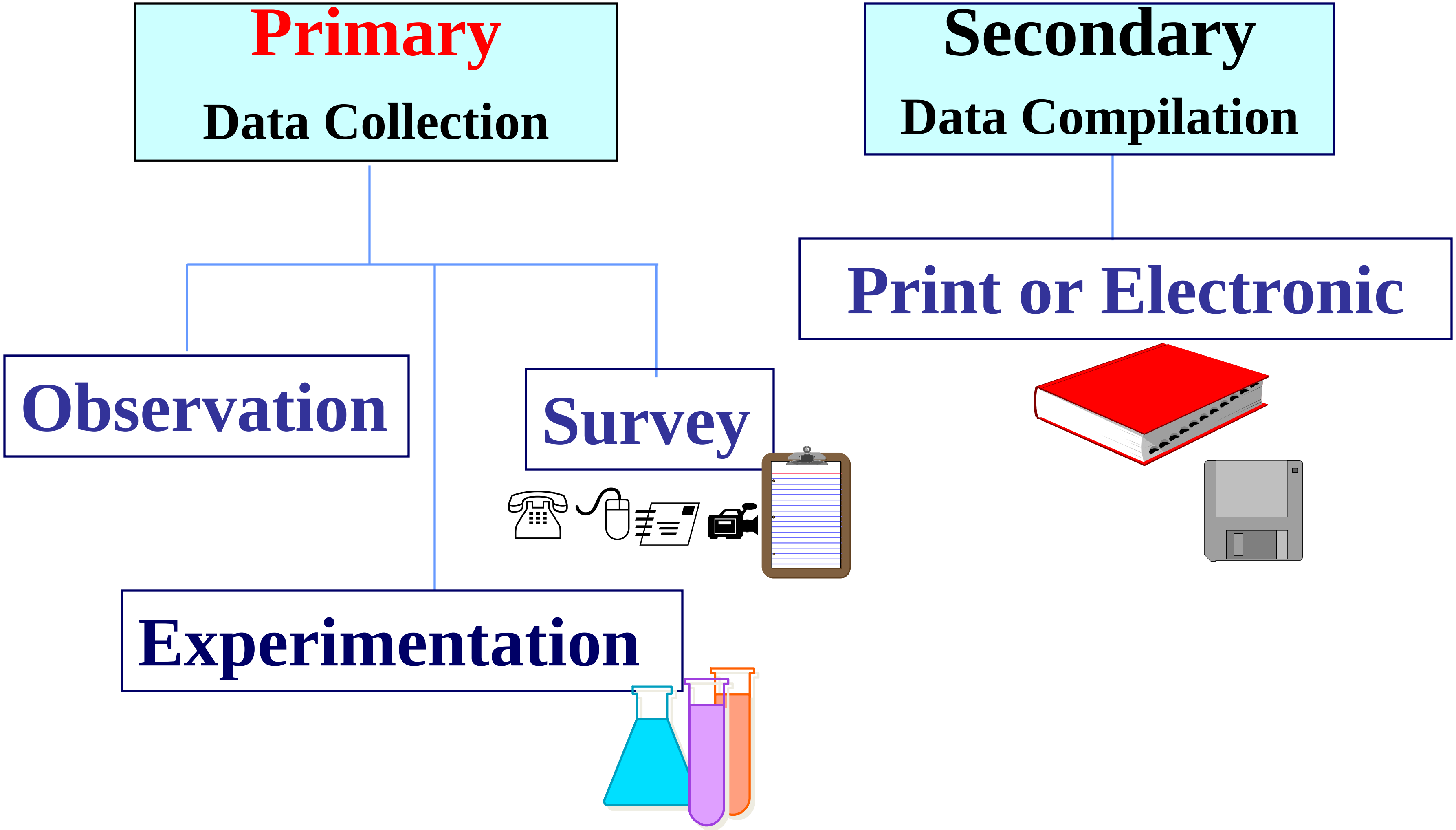
- Collect data
  - e.g. Survey

- Characterize data
  - e.g. Sample mean = $\frac{\sum X_i}{n}$
- Present data
  - e.g. Tables and graphs

Why We Need Data

- To provide input to survey
- To provide input to study
- To measure performance of service or production process
- To evaluate conformance to standards
- To assist in formulating alternative courses of action
- To satisfy curiosity

Data Sources



Primary

Data Collection

Secondary

Data Compilation

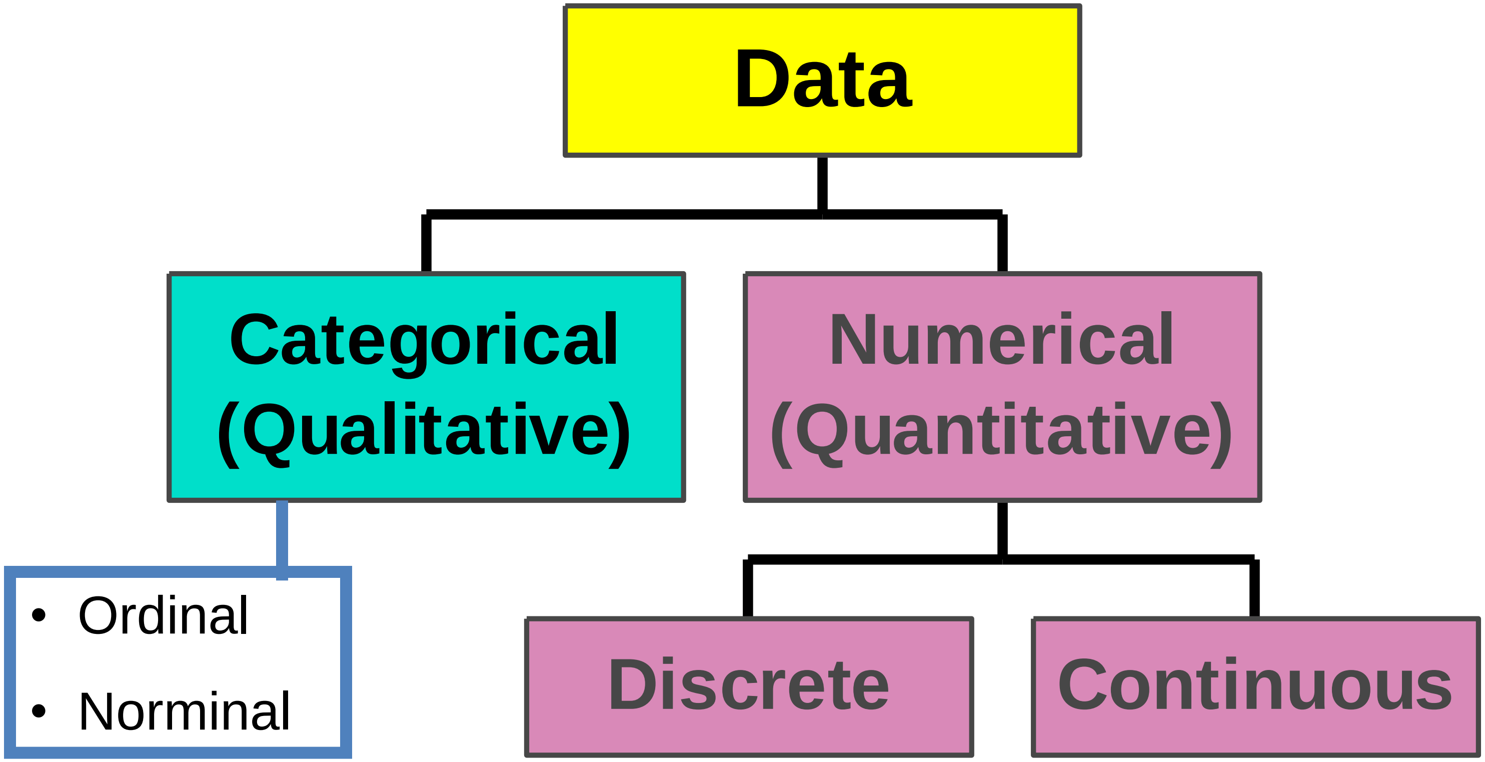
Observation

Survey

Experimentation

Print or Electronic

Types of Data



Data

Categorical (Qualitative)

Numerical (Quantitative)

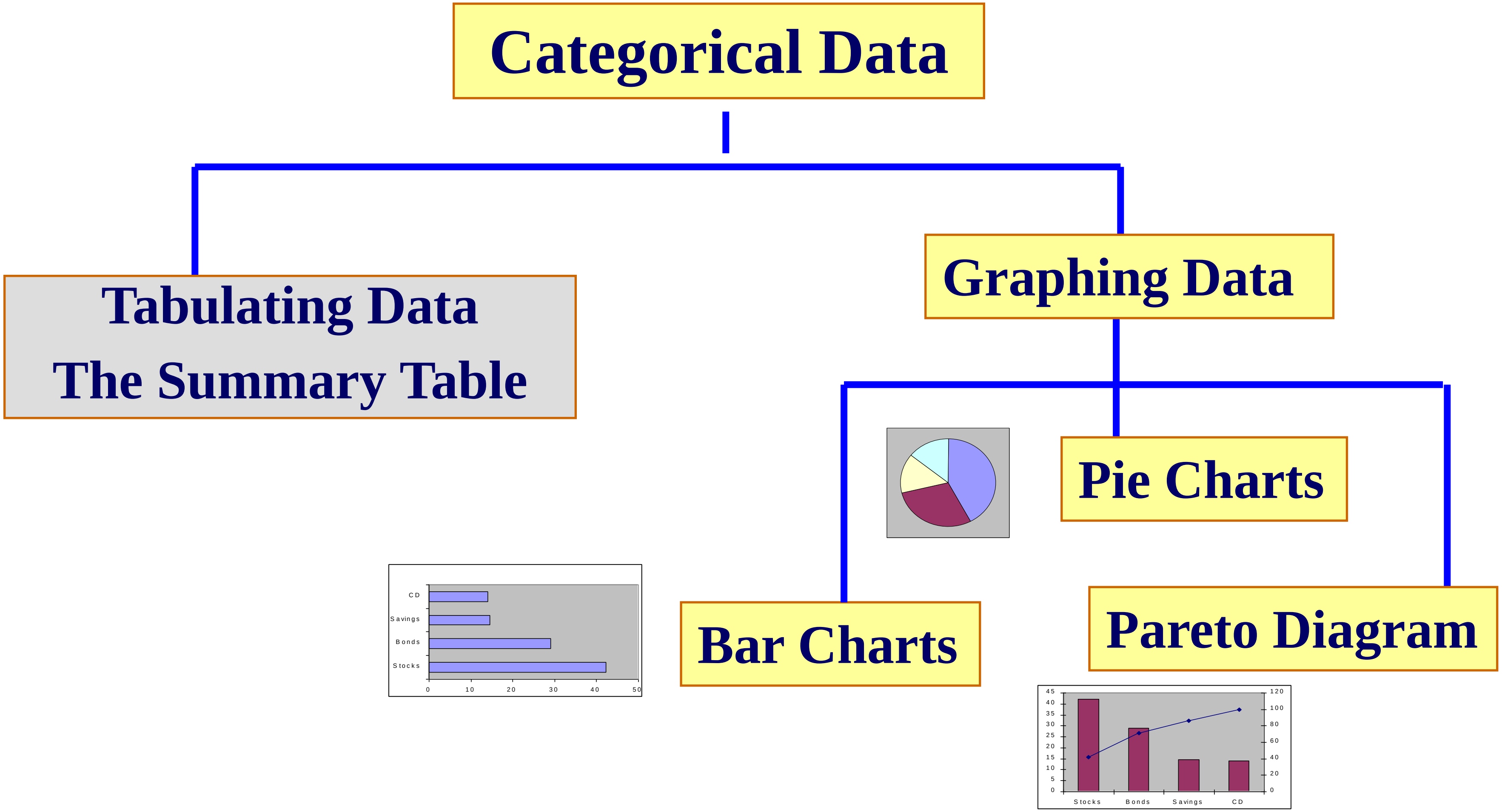
- Ordinal
- Norminal

Discrete

Continuous

Statistics for manager using Excel, 3e © 2000 Prentice-Hall, Inc.

Graphing Categorical Data: Univariate Data



Categorical Data

Tabulating Data

The Summary Table

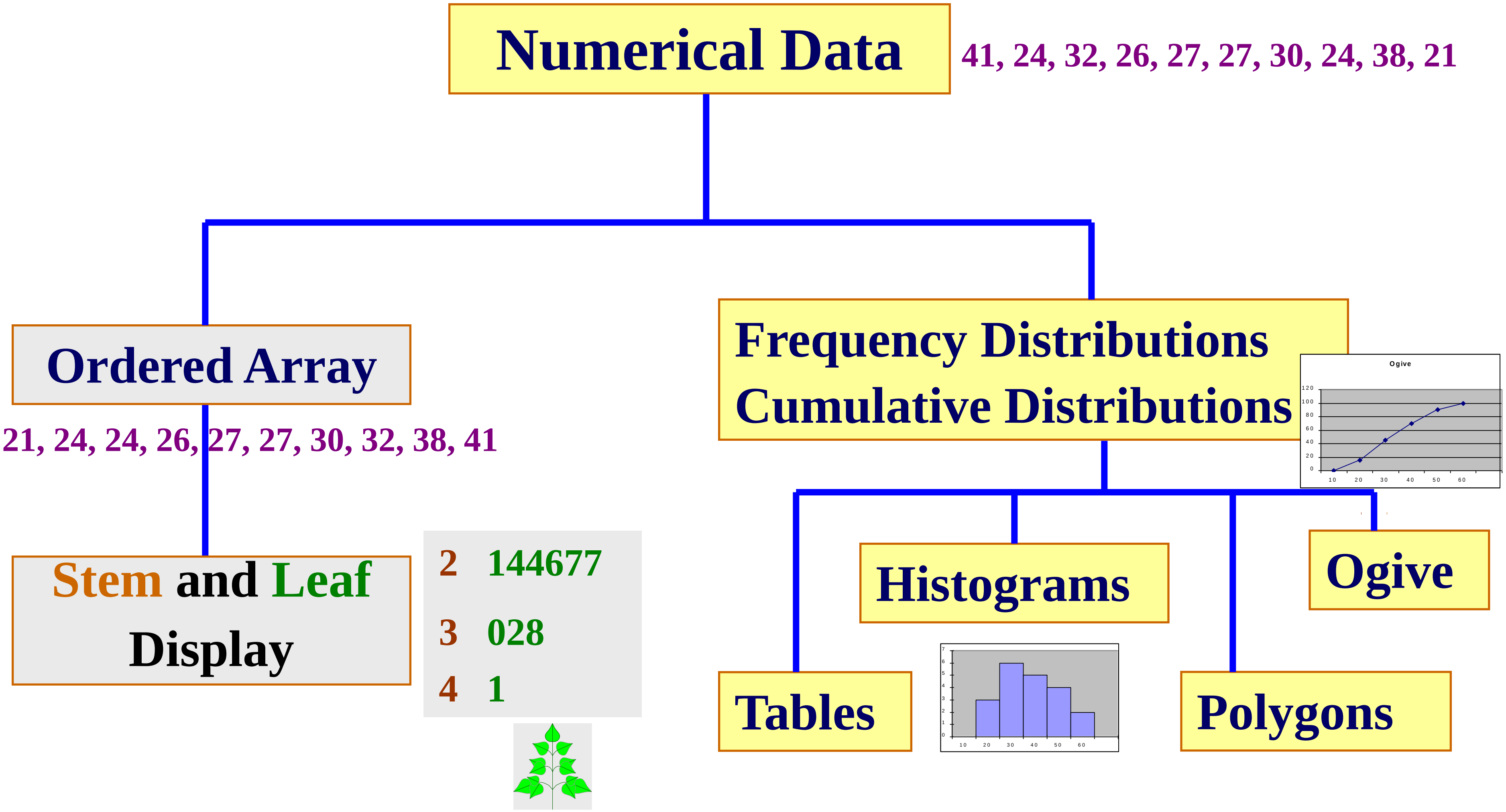
Graphing Data

Pie Charts

Bar Charts

Pareto Diagram

Tabulating and Graphing Numerical Data



Numerical Data

41, 24, 32, 26, 27, 27, 30, 24, 38, 21

Ordered Array

21, 24, 24, 26, 27, 27, 30, 32, 38, 41

Stem and Leaf Display

2	144677
3	028
4	1

Frequency Distributions

Cumulative Distributions

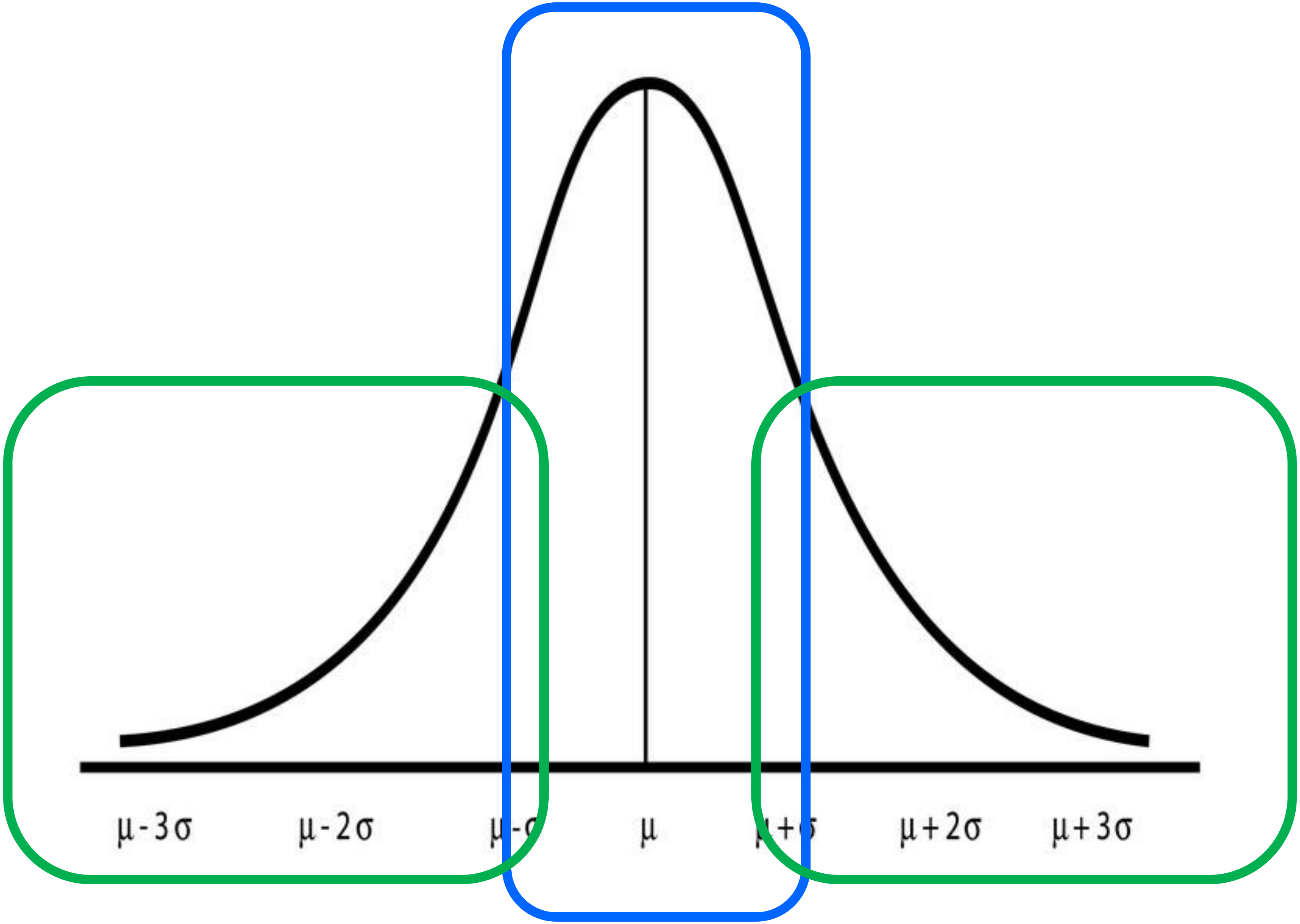
Histograms

Ogive

Tables

Polygons

- Measures of central tendency
  - Mean
  - Median
  - Mode
- Measures of dispersion
  - Range
  - Interquartile range
  - Variance and standard deviation
  - Coefficient of variation

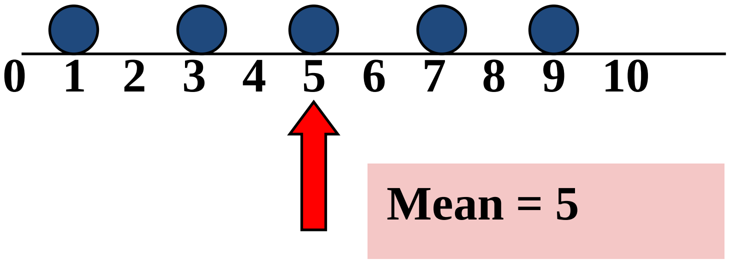

μ-3σ
μ-2σ
μ-σ
μ
μ+σ
μ+2σ
μ+3σ

Mean

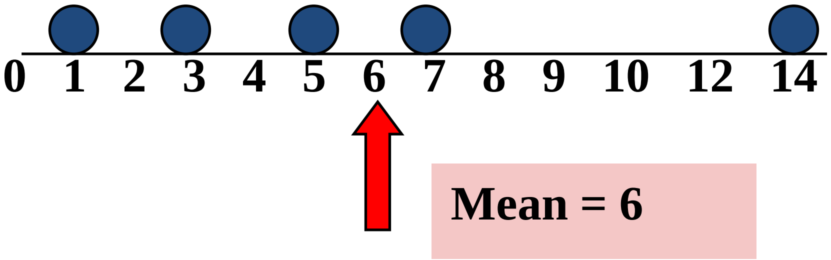
- The sum of the measurements divided by the total number of measurements or better known as the average.

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

- There is only 1 mean.
- Work well if data is reasonably symmetric and unimodal (“bell-shaped”)
- Not good measurement if got extreme data values (“outliers”) are present & data are skewed
- Value is influences by extreme measurements
- Applicable to quantitative data only.



Mean = 5



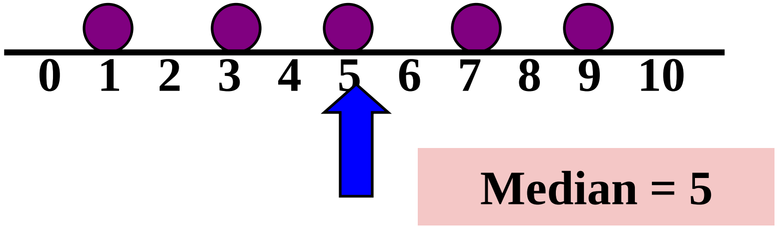
Mean = 6

$$\bar{x} = \frac{1+3+5+7+9}{5} = \frac{25}{5}$$

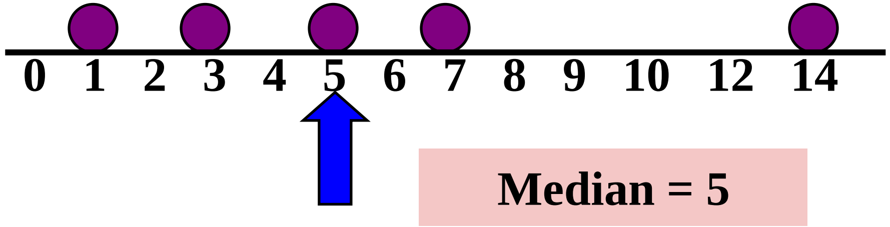
$$\bar{x} = \frac{1+3+5+7+14}{5} = \frac{30}{5}$$

Median

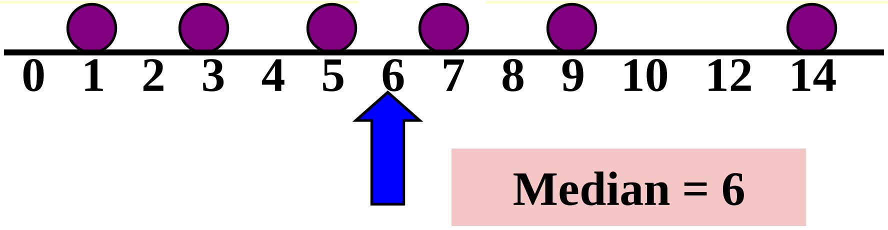
- The middle value when the measurements are arranged from lowest to highest.
- 50% of the measurement lie above it and 50% fall below it.
- Often used to measure the midpoint of a large set of measurement.
  - median = 50th percentile = second quartile (Q_2)
- There is only 1 median
- Not influenced by extreme measurements.
- Applicable to quantitative data only.



Median = 5



Median = 5

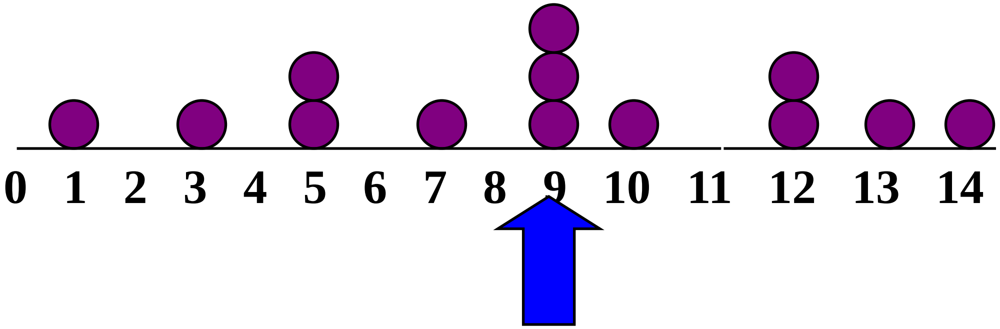


Median = 6

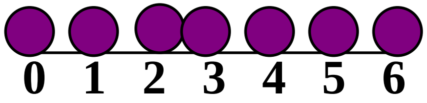
- If the number of observations (*n*) is odd, the median is the middle value, or the [(*n*+1)/2]th observation.
- If *n* is even, the median is usually calculated as the average of the two middlemost values- that is, the average of the [(*n*/2)]th observation and the [(*n*/2) + 1]th observation.

Mode

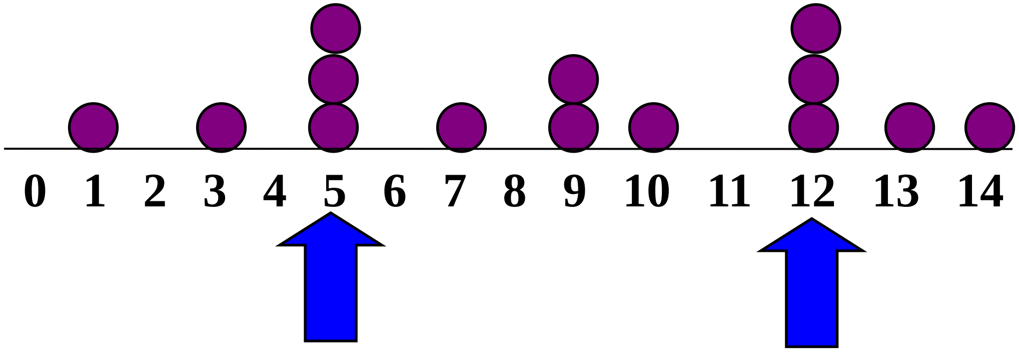
- The measurement that occurs most often (with the highest frequency)
- Commonly used as a measure of popularity.
- There can be more than 1 mode.
  - Unimodal, bimodal or multimodal
- Not influence by extreme measurements.
- Applicable for both qualitative and quantitative data.



Mode = 9



No Mode



Mode = 5, 12

Measures of Variability

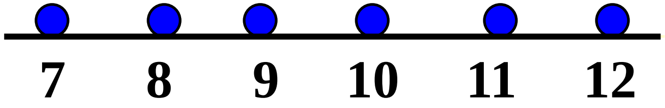
- It is not sufficient to describe a data set using only measures of central tendency
- Need to determine how dispersed/ spread out the data is.
- Measures of variability/spread includes
  - Range
  - Percentile / Quartile
  - Deviation / Standard Deviation (sisihan piawai)
  - Variance
  - Coefficient of variation

Range

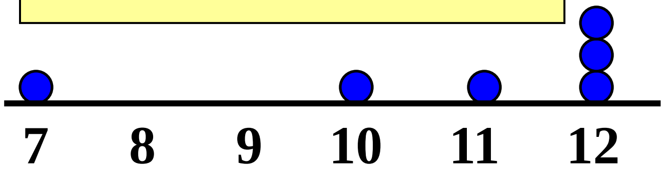
- Measure of variation
- The difference between the largest and the smallest measurement / observations of the set.
- It is easy to compute but very sensitive to outliers.
- Does not give much information about the pattern of variability
- Ignores the way in which data are distributed

$$\text{Range} = X_{\text{Largest}} - X_{\text{Smallest}}$$

Range = 12 - 7 = 5



Range = 12 - 7 = 5

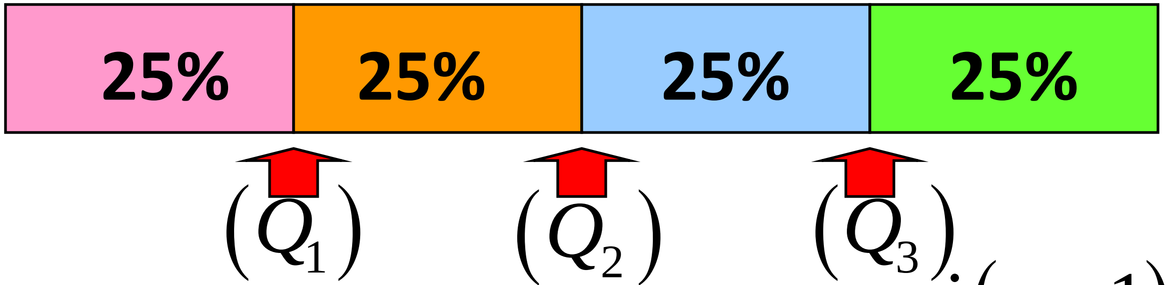


Percentile / Quartile

- The pth percentile of a set of n measurements arranged in order of magnitude is that value that has at most p% of the measurements below it and at most (100 – p) % above it.
- Example: 60th percentile has 60% of the data below it and 40% above it.
- Percentile of interest are the 25th, 50th, 75th, percentiles often called the lower quartile, median, and upper quartile.
- Interquartile range – difference between the upper and lower quartile

Quartiles

- Split Ordered Data into 4 Quarters



- Position of i-th Quartile $(Q_i) = \frac{i(n+1)}{4}$

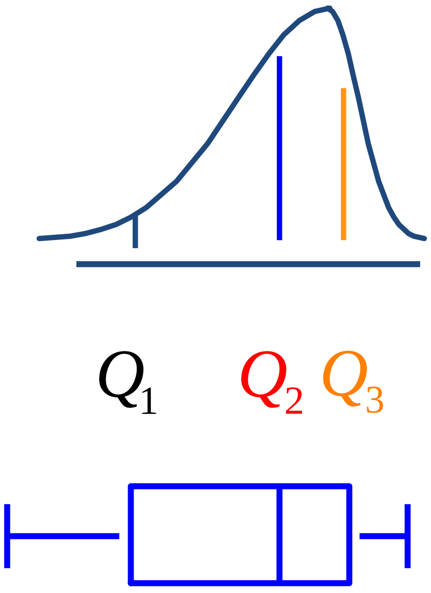
Data in Ordered Array: 11 12 13 16 16 17 18 21 22

$$\text{Position of } Q_1 = \frac{1(9+1)}{4} = 2.5 \qquad Q_1 = \frac{(12+13)}{2} = 12.5$$

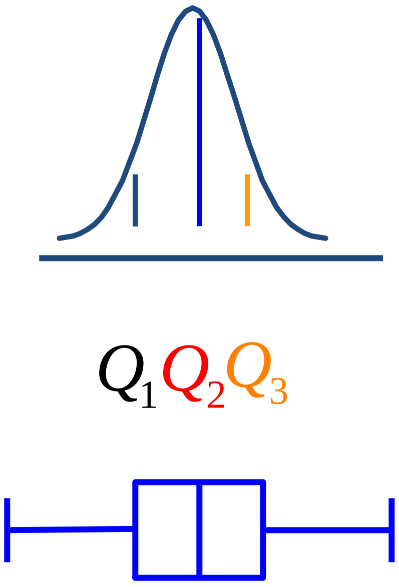
- Q_1 and Q_3 are Measures of Noncentral Location
- Q_2 = Median, A Measure of Central Tendency

Distribution Shape and Box-and-Whisker Plot

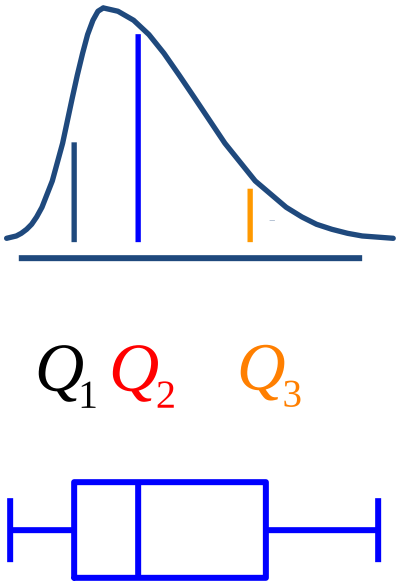
Left-Skewed



Symmetric



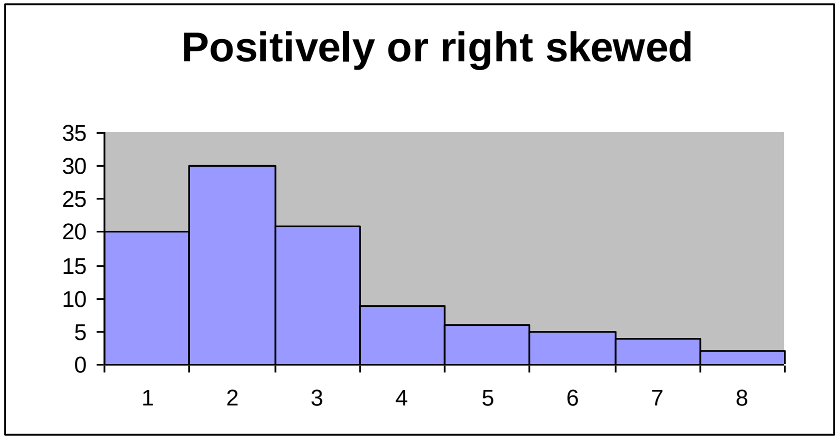
Right-Skewed



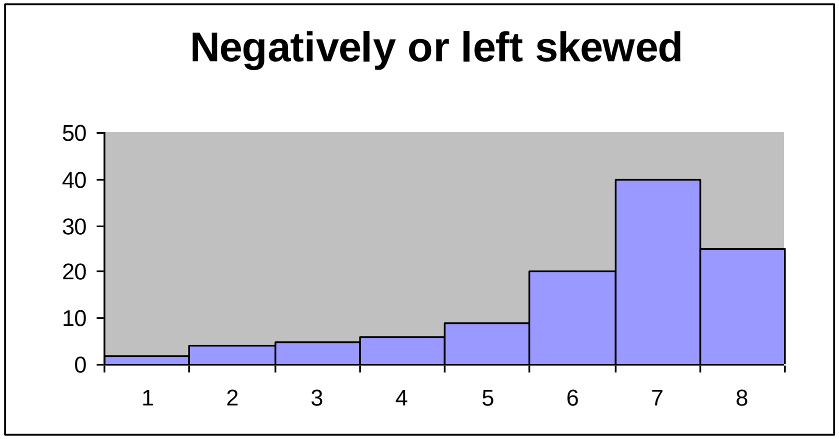
Skewness

- Relationship of the mode, median, mean and trimmed mean is reflected through the skewness of the data.
- Skewness of the data measures how the data is distributed.
- Zero Skewness
  - symmetrical (Mode = Median = Mean)
- Positive Skewness
  - skewed to the right (Mode < Median < Mean)
- Negative Skewness
  - skewed to the left (Mode > Median > Mean)

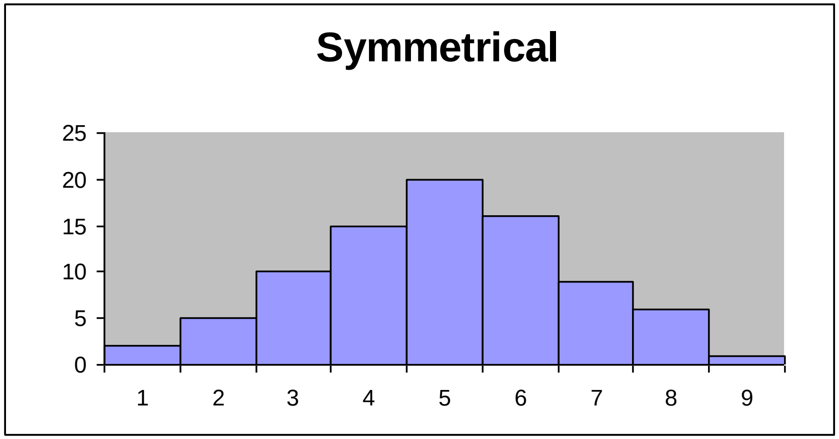
Skewness



Mode<Median<Mean



Mean<Median<Mode



Mode=Median=Mean

Variance and Standard Deviation

- The variance of a set of n measurements x_1, x_2, … , x_n with mean y is the sum of the squared deviations divided by n – 1.

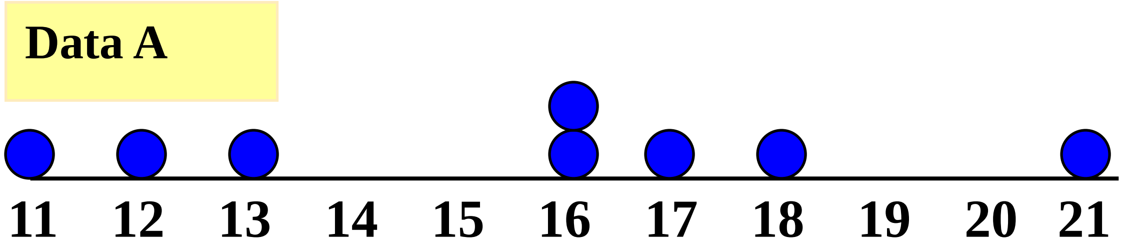
$$\sigma^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$$

- The standard deviation of a set of measurement is defined to be the positive square root of the variance.

$$\sigma = \sqrt{\sigma^2}$$

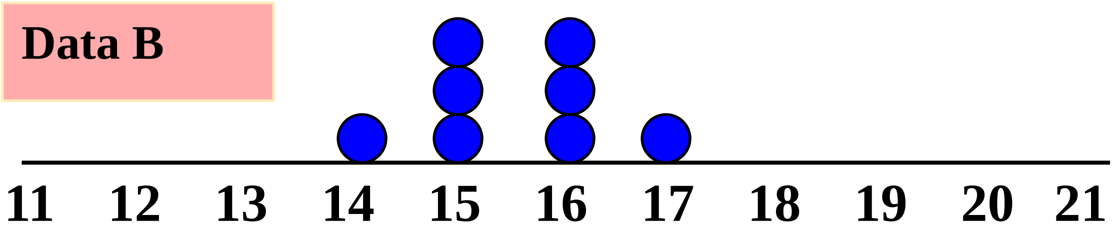
- Both measure how spread out the data is from the mean.

Comparing Standard Deviations



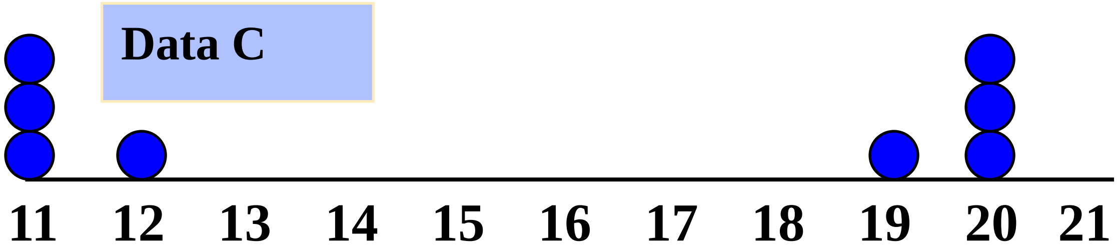
Mean = 15.5

S = 3.338



Mean = 15.5

S = .9258



Mean = 15.5

S = 4.57

Coefficient of Variation

- Measures the variability in the values in a population relative to the magnitude of the population mean.
- $$CV = \frac{\text{Standard Deviation}}{|\text{Mean}|}$$
- The CV is a unit-free number, it is useful when comparing variations of different sets of data.

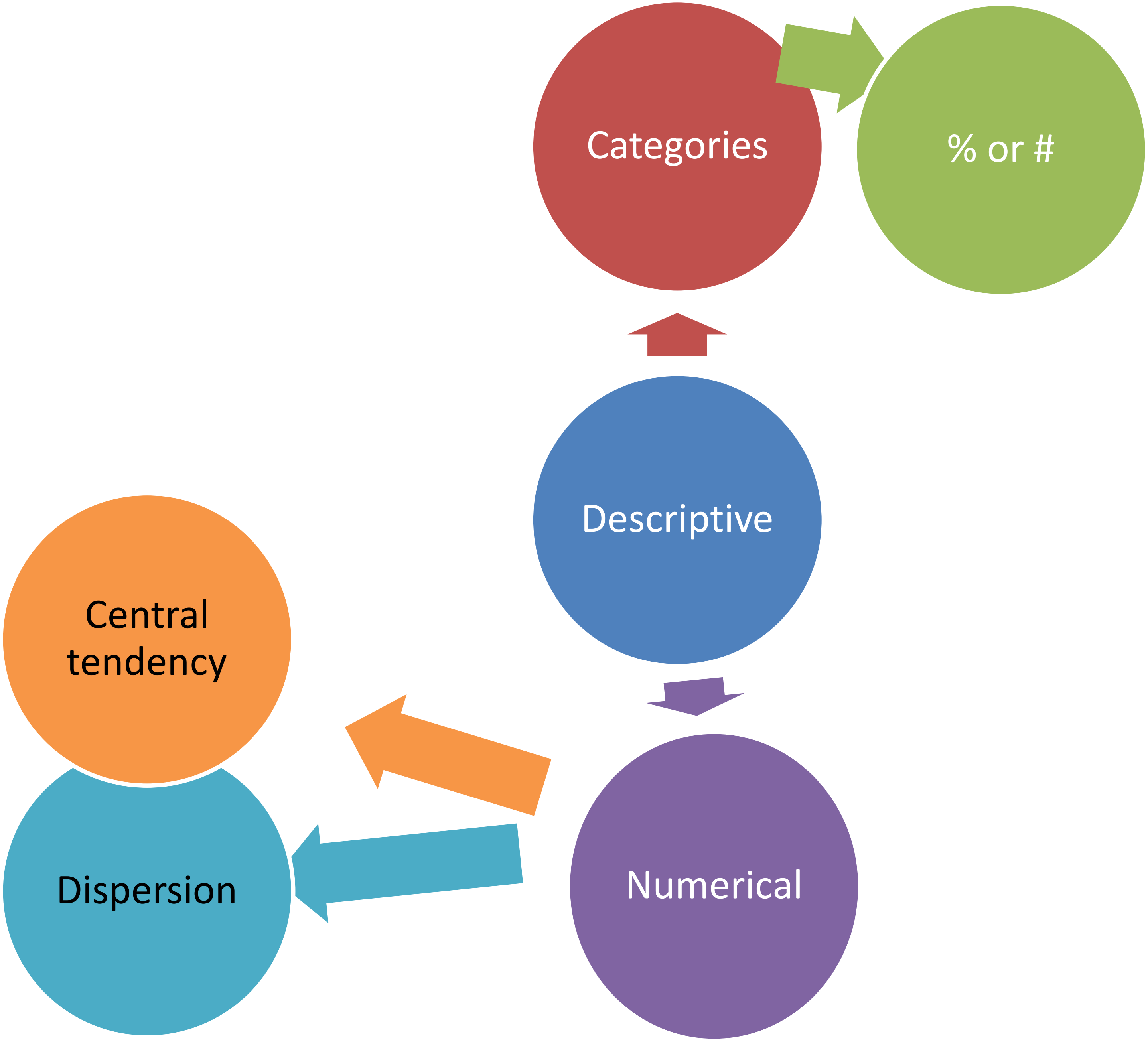
Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5
- Stock B:
  - Average price last year = $100
  - Standard deviation = $5
- Coefficient of variation:
  - Stock A:

$$CV = \left(\frac{S}{\overline{X}} \right) 100\% = \left(\frac{\$5}{\$50} \right) 100\% = 10\%$$

  - Stock B:

$$CV = \left(\frac{S}{\overline{X}} \right) 100\% = \left(\frac{\$5}{\$100} \right) 100\% = 5\%$$



Categories

% or #

Descriptive

Central tendency

Dispersion

Numerical

Thank you