INTRODUCTION TO DATA ANALYTICS

Basic Statistics

Dr. Rathachai Chawuthai

Department of Computer Engineering
Faculty of Engineering
King Mongkut's Institute of Technology Ladkrabang

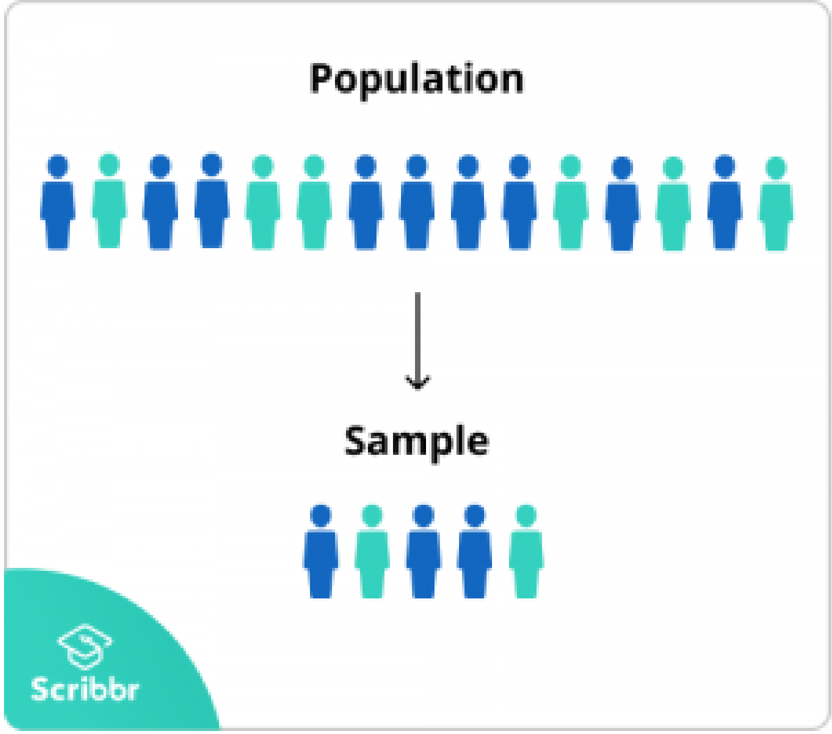
Agenda

- Introduction
- Data Types
- Central Tendency
- Measure of Dispersion
- Distribution
- Probability
- Correlation

Introduction

Population vs sample

- The **population** is the entire group that you want to draw conclusions about.
- The **sample** is the specific group of individuals that you will collect data from.



Ref: • https://www.scribbr.com/methodology/sampling-methods/

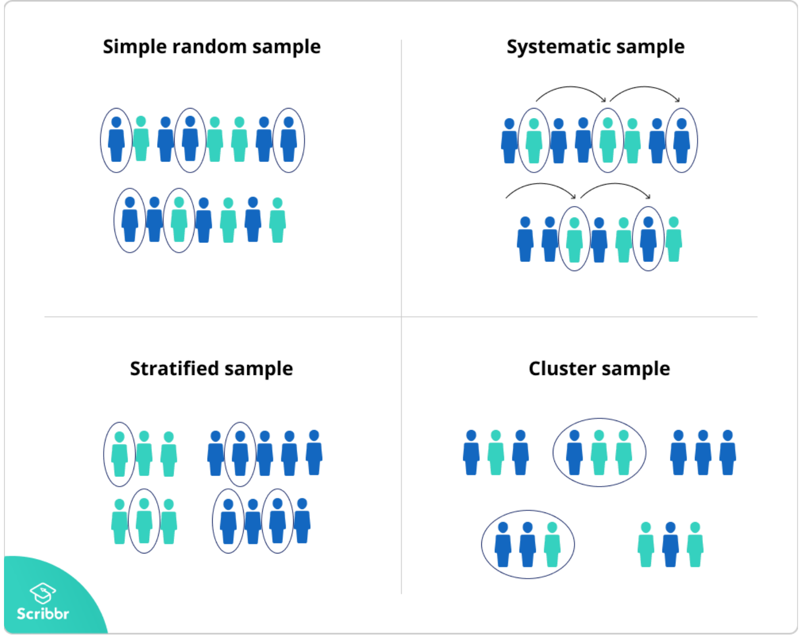
Sampling Methods

- When you conduct research about a group of people, it's rarely possible to collect data from every person in that group. Instead, you select a sample. The sample is the group of individuals who will actually participate in the research.
- To draw valid conclusions from your results, you have to carefully decide how you will select a sample that is representative of the group as a whole. There are two types of sampling methods:
  - **Probability sampling** involves random selection, allowing you to make statistical inferences about the whole group.
  - **Non-probability sampling** involves non-random selection based on convenience or other criteria, allowing you to easily collect initial data.

Ref: • (content) https://www.scribbr.com/methodology/sampling-methods/

Probability sampling methods

- Probability sampling means that every member of the population has a chance of being selected. If you want to produce results that are representative of the whole population, you need to use a probability sampling technique.
- There are four main types of probability sample.



Ref: • (content) https://www.scribbr.com/methodology/sampling-methods/

Simple random sampling

- In a simple random sample, every member of the population has an equal chance of being selected. Your sampling frame should include the whole population.
- To conduct this type of sampling, you can use tools like random number generators or other techniques that are based entirely on chance.
- **Example**
  - You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

Ref: • (content) https://www.scribbr.com/methodology/sampling-methods/

Systematic sampling

- Systematic sampling is similar to simple random sampling, but it is usually slightly easier to conduct. Every member of the population is listed with a number, but instead of randomly generating numbers, individuals are chosen at regular intervals.
- **Example**
  - All employees of the company are listed in alphabetical order. From the first 10 numbers, you randomly select a starting point: number 6. From number 6 onwards, every 10th person on the list is selected (6, 16, 26, 36, and so on), and you end up with a sample of 100 people.
- If you use this technique, it is important to make sure that there is no hidden pattern in the list that might skew the sample. For example, if the HR database groups employees by team, and team members are listed in order of seniority, there is a risk that your interval might skip over people in junior roles, resulting in a sample that is skewed towards senior employees.

Ref: • (content) https://www.scribbr.com/methodology/sampling-methods/

Stratified sampling

- This sampling method is appropriate when the population has mixed characteristics, and you want to ensure that every characteristic is proportionally represented in the sample.
- You divide the population into subgroups (called strata) based on the relevant characteristic (e.g. gender, age range, income bracket, job role).
- From the overall proportions of the population, you calculate how many people should be sampled from each subgroup. Then you use random or systematic sampling to select a sample from each subgroup.
- **Example**
  - The company has 800 female employees and 200 male employees. You want to ensure that the sample reflects the gender balance of the company, so you sort the population into two strata based on gender. Then you use random sampling on each group, selecting 80 women and 20 men, which gives you a representative sample of 100 people.

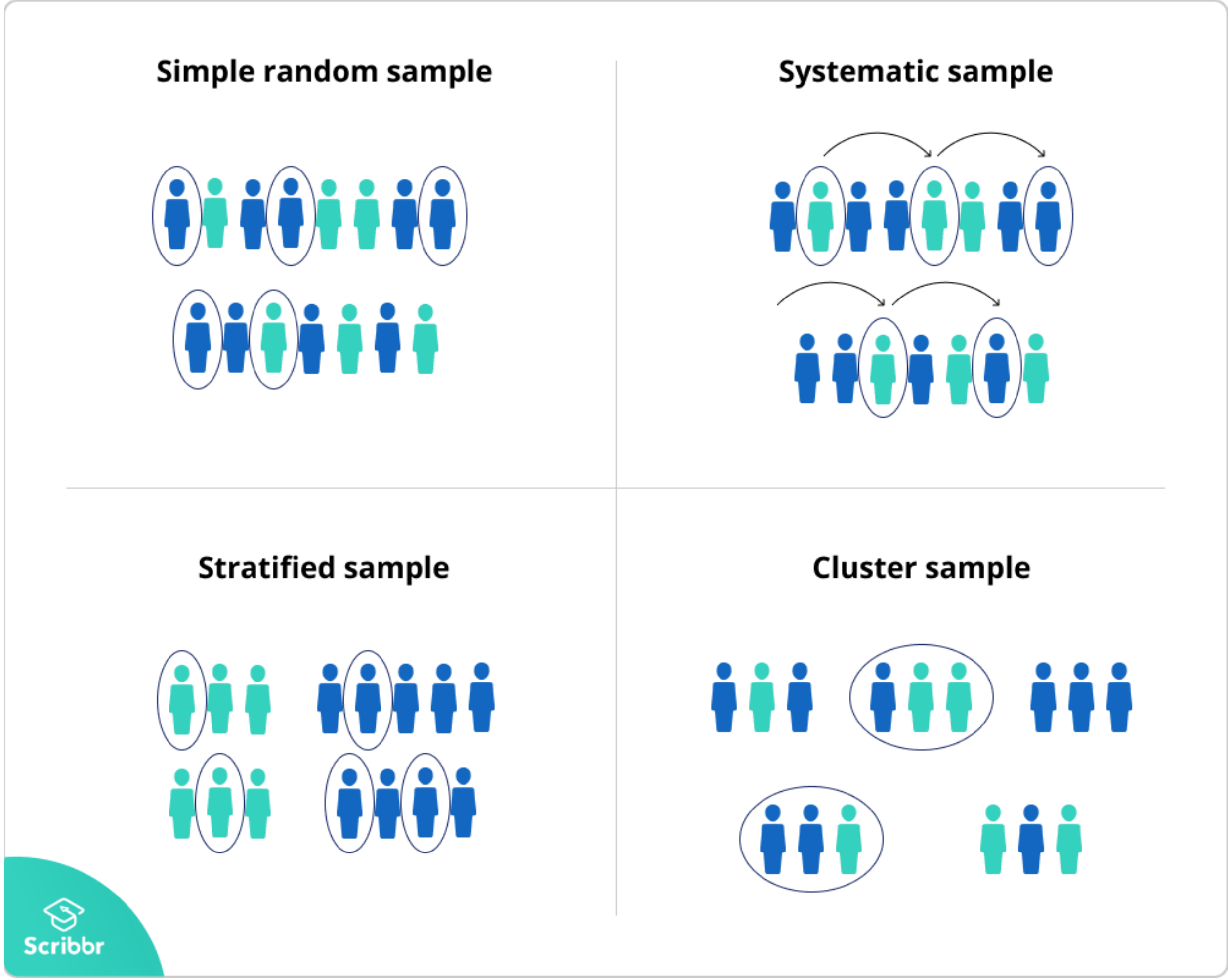
Ref: • (content) https://www.scribbr.com/methodology/sampling-methods/

Cluster sampling

- Cluster sampling also involves dividing the population into subgroups, but each subgroup should have similar characteristics to the whole sample. Instead of sampling individuals from each subgroup, you randomly select entire subgroups.
- If it is practically possible, you might include every individual from each sampled cluster. If the clusters themselves are large, you can also sample individuals from within each cluster using one of the techniques above.
- This method is good for dealing with large and dispersed populations, but there is more risk of error in the sample, as there could be substantial differences between clusters. It's difficult to guarantee that the sampled clusters are really representative of the whole population.
- Example
  - The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters.

Ref: • (content) https://www.scribbr.com/methodology/sampling-methods/

Probability sampling methods: Summary



Ref: • (image) https://www.scribbr.com/methodology/sampling-methods/

Convenience sampling

- Convenience sampling is perhaps the easiest method of sampling, because participants are selected based on availability and willingness to take part. Useful results can be obtained, but the results are prone to significant bias, because those who volunteer to take part may be different from those who choose not to (volunteer bias), and the sample may not be representative of other characteristics, such as age or sex.
- Note: volunteer bias is a risk of all non-probability sampling methods.

Ref: • (content) https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population

Voluntary response sampling

- Similar to a convenience sample, a voluntary response sample is mainly based on ease of access. Instead of the researcher choosing participants and directly contacting them, people volunteer themselves (e.g. by responding to a public online survey).
- Voluntary response samples are always at least somewhat biased, as some people will inherently be more likely to volunteer than others.
- **Example**
  - You send out the survey to all students at your university and a lot of students decide to complete it. This can certainly give you some insight into the topic, but the people who responded are more likely to be those who have strong opinions about the student support services, so you can’t be sure that their opinions are representative of all students.

Ref: • (content) https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population

Purposive Sampling

- Also known as selective, or subjective, sampling, this technique relies on the judgement of the researcher when choosing who to ask to participate. Researchers may implicitly thus choose a "representative" sample to suit their needs, or specifically approach individuals with certain characteristics. This approach is often used by the media when canvassing the public for opinions and in qualitative research.
- Judgement sampling has the advantage of being time-and cost-effective to perform whilst resulting in a range of responses (particularly useful in qualitative research). However, in addition to volunteer bias, it is also prone to errors of judgement by the researcher and the findings, whilst being potentially broad, will not necessarily be representative.

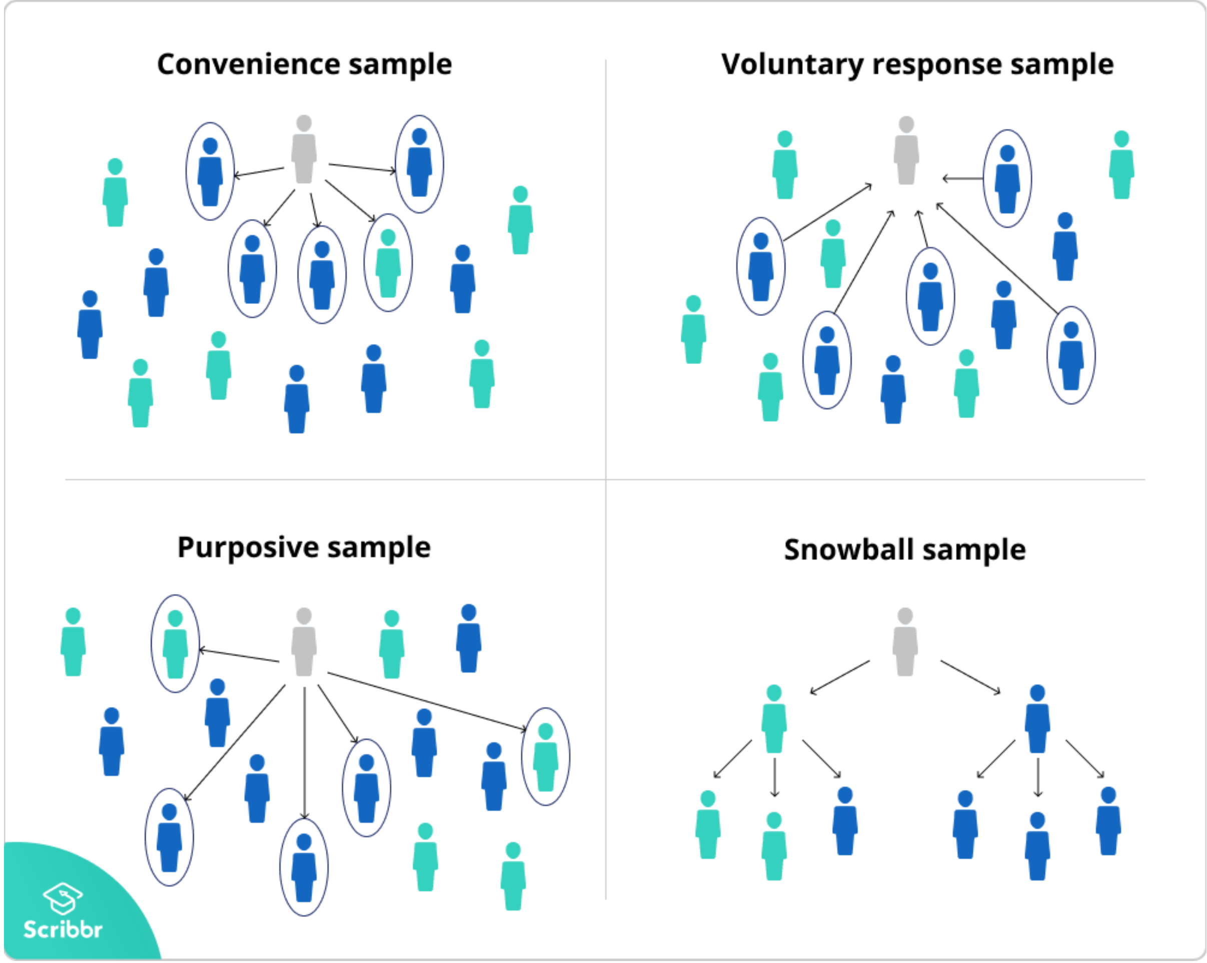
Ref: • (content) https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population

Snowball sampling

- This method is commonly used in social sciences when investigating hard-to-reach groups. Existing subjects are asked to nominate further subjects known to them, so the sample increases in size like a rolling snowball. For example, when carrying out a survey of risk behaviors amongst intravenous drug users, participants may be asked to nominate other users to be interviewed.
- Snowball sampling can be effective when a sampling frame is difficult to identify. However, by selecting friends and acquaintances of subjects already investigated, there is a significant risk of selection bias (choosing a large number of people with similar characteristics or views to the initial individual identified).

Ref: • (content) https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/methods-of-sampling-population

Non-Probability sampling methods: Summary



Ref:
- (image) https://www.scribbr.com/methodology/sampling-methods/

Bias in sampling

- There are five important potential sources of bias that should be considered when selecting a sample, irrespective of the method used. Sampling bias may be introduced when:

1. Any pre-agreed sampling rules are deviated from
2. People in hard-to-reach groups are omitted
3. Selected individuals are replaced with others, for example if they are difficult to contact
4. There are low response rates
5. An out-of-date list is used as the sample frame (for example, if it excludes people who have recently moved to an area)

Ref: • (content) https://www.scribbr.com/methodology/sampling-methods/

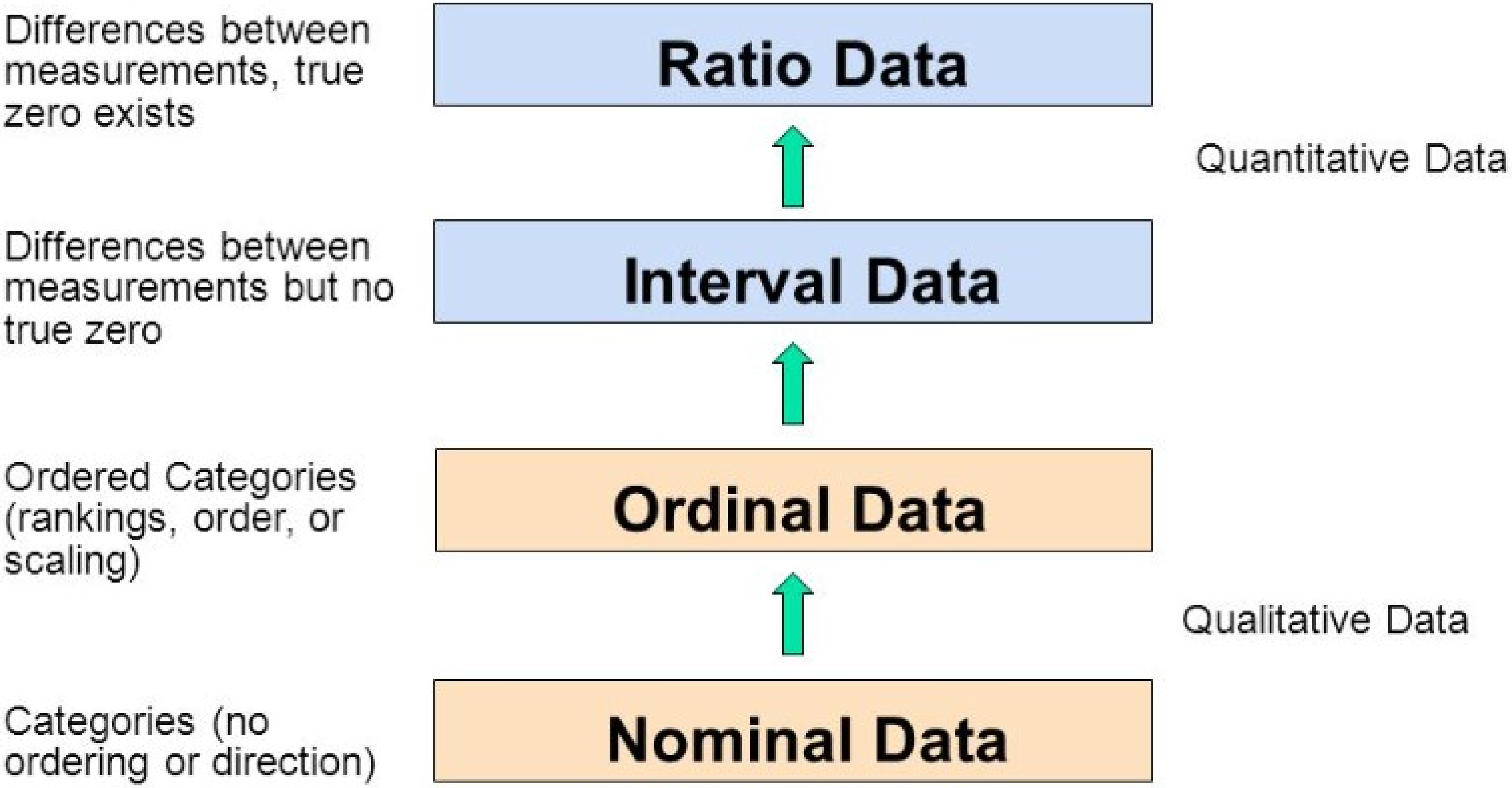
Data Types

Statistics

- Statistics is the discipline that concerns the collection, organization, displaying, analysis, interpretation and presentation of data.
- The main difference between a population and sample has to do with how observations are assigned to the data set.
  - A population includes all of the elements from a set of data.
  - A sample consists one or more observations drawn from the population.
- Data
  - Quantitative Data : continuous data, discrete data
  - Qualitative Data: categorical data

Ref:
- (content) https://en.wikipedia.org/wiki/Statistics
- (content) https://stattrek.com/sampling/populations-and-samples.aspx

Scale of Data



Differences between measurements, true zero exists

Ratio Data

Quantitative Data

Differences between measurements but no true zero

Interval Data

Ordered Categories (rankings, order, or scaling)

Ordinal Data

Qualitative Data

Categories (no ordering or direction)

Nominal Data

Ref: • (image) https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

Scale of Data

- **Nominal** scale
  - Categorical data without order e.g. gender, job, department
- **Ordinal** scale
  - Categorical data with order e.g. grade, rating
- **Interval** scale
  - Ordinal scale (zero has meaning) e.g. temperature in C vs F
- **Ratio** scale
  - Interval scale (zero is zero) e.g. income, height

Nominal

- A nominal scale describes a variable with categories that do not have a natural order or ranking. You can code nominal variables with numbers if you want, but the order is arbitrary and any calculations, such as computing a mean, median, or standard deviation, would be meaningless.
- Examples of nominal variables include:
  - genotype, blood type,
  - zip code, gender,
  - race, eye color,
  - political party

Ref: • (content) https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

Ordinal

- An ordinal scale is one where the order matters but not the difference between values.
- Examples of ordinal variables include:
  - socio economic status (“low income”,”middle income”,”high income”),
  - education level (“high school”,”BS”,”MS”,”PhD”),
  - income level (“less than 50K”, “50K-100K”, “over 100K”),
  - satisfaction rating (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”).
- Note the differences between adjacent categories do not necessarily have the same meaning. For example, the difference between the two income levels “less than 50K” and “50K-100K” does not have the same meaning as the difference between the two income levels “50K-100K” and “over 100K”.

Ref: • (content) https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

Interval

- An interval scale is one where there is order and the difference between two values is meaningful.
- Examples of interval variables include:
  - temperature (Farenheit),
  - temperature (Celcius),
  - pH

Ref: • (content) https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

Ratio

- A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.
- Examples of ratio variables include:
  - enzyme activity, dose amount, reaction rate, flow rate, concentration, pulse, weight, length, temperature in Kelvin (0.0 Kelvin really does mean “no heat”), survival time.
- When working with ratio variables, but not interval variables, the ratio of two measurements has a meaningful interpretation. For example, because weight is a ratio variable, a weight of 4 grams is twice as heavy as a weight of 2 grams. However, a temperature of 10 degrees C should not be considered twice as hot as 5 degrees C. If it were, a conflict would be created because 10 degrees C is 50 degrees F and 5 degrees C is 41 degrees F. Clearly, 50 degrees is not twice 41 degrees. Another example, a pH of 3 is not twice as acidic as a pH of 6, because pH is not a ratio variable.

Ref: • (content) https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

Measurement

Provides:	Nominal	Ordinal	Interval	Ratio
The “order” of value is known		✓	✓	✓
“Counts”, aka “Frequency of Distribution	✓	✓	✓	✓
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓
Can quantify the difference between each value			✓	✓
Can add or subtract values			✓	✓
Can multiply and divide values				✓
Has “true zero”				✓

Ref: • (content) https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

Binary

Possible values	0, 1 (arbitrary labels)
Example usage	binary outcome ("yes/no", "true/false", "success/failure", etc.)
Permissible statistics	mode
Comparison	No

Ref: • (content) https://en.wikipedia.org/wiki/Statistical_data_type

Categorical

Possible values	1, 2, ..., K (arbitrary labels)
Example usage	categorical outcome (specific blood type, political party, word, etc.)
Permissible statistics	mode, Chi-squared
Comparison	No

Ref: • (content) https://en.wikipedia.org/wiki/Statistical_data_type

Natural Number

Possible values	nonnegative integers (0, 1, ...)
Example usage	number of items (telephone calls, people, molecules, births, deaths, etc.) in given interval/area/volume
Permissible statistics	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation
Comparison	Yes

Ref:

- (content) https://en.wikipedia.org/wiki/Statistical_data_type

Real number

Possible values	real number
Example usage	temperature, relative distance, location parameter, etc. (or approximately, anything not varying over a large scale)
Permissible statistics	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation
Comparison	Yes
Operation	round(3.14) = 3 round(3.14, 1) = 3.1

Ref: • (content) https://en.wikipedia.org/wiki/Statistical_data_type

Date Time

Possible values	2009-06-15 13:45:30 ('YYYY-MM-DD HH:MM:SS’)
Example usage	Date Time
Permissible statistics	Mean, Mode, Median (convert to timestamp first)
Comparison	Yes
Operation	Get Year, Get Month, Get Day, Get Hour, Get Minute, Get Second, Get Timespan

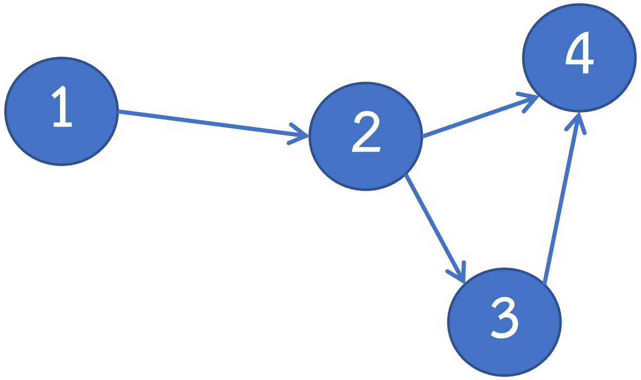
Latitude/Longitude

Possible values	35.929673,-78.948237
Example usage	Location
Permissible statistics	Mean, Mode, Median
Comparison	Yes
Operation	Distance between two location

String

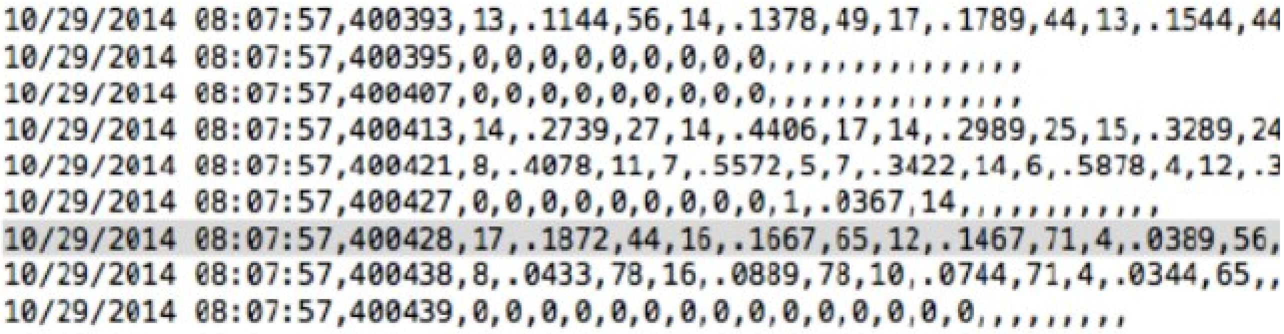
Possible values	“Hello”, “World”, “Hello World”
Example usage	Name, Label
Permissible statistics	Mode
Comparison	by alphabet or length
Operation	Concatenate, split, Upper, Lower, Find, Not

Graph

Possible values	(N1, N2), (N2, N3), (N2, N4), (N3, N4) 
Example usage	Social Network
Permissible statistics	Average Degree
Comparison	no

Central Tendency

Sample



```
10/29/2014 08:07:57,400393,13,.1144,56,14,.1378,49,17,.1789,44,13,.1544,44
10/29/2014 08:07:57,400395,0,0,0,0,0,0,0,0,0,,,,,,,,,,,,,,,
10/29/2014 08:07:57,400407,0,0,0,0,0,0,0,0,0,,,,,,,,,,,,,,,
10/29/2014 08:07:57,400413,14,.2739,27,14,.4406,17,14,.2989,25,15,.3289,24
10/29/2014 08:07:57,400421,8,.4078,11,7,.5572,5,7,.3422,14,6,.5878,4,12,.3
10/29/2014 08:07:57,400427,0,0,0,0,0,0,0,0,1,.0367,14,,,,,,,,,,,,
10/29/2014 08:07:57,400428,17,.1872,44,16,.1667,65,12,.1467,71,4,.0389,56,
10/29/2014 08:07:57,400438,8,.0433,78,16,.0889,78,10,.0744,71,4,.0344,65,,
10/29/2014 08:07:57,400439,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,,,,,,,,,
```

sample

sample

sample

sample

#	Body Temp	Headache	Nausea	Flu
1	high	yes	-	yes
2	very high	yes	yes	yes
3	normal	-	-	-
4	high	yes	yes	yes
5	high	-	yes	-

sample

sample

sample

sample

sample

Arithmetic Mean

- The most common type of average is the arithmetic mean. If n numbers are given, each number denoted by ai (where i = 1,2, …, n), the arithmetic mean is the sum of the as divided by n or

$$AM = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n}\left(a_1 + a_2 + \cdots + a_n\right)$$

Ref: • https://en.wikipedia.org/wiki/Average

Harmonic Mean

- Harmonic mean for a non-empty collection of numbers a1, a2, …, an, all different from 0, is defined as the reciprocal of the arithmetic mean of the reciprocals of the ai's:

$$HM = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}} = \frac{n}{\frac{1}{a_1}+\frac{1}{a_2}+\cdots+\frac{1}{a_n}}$$

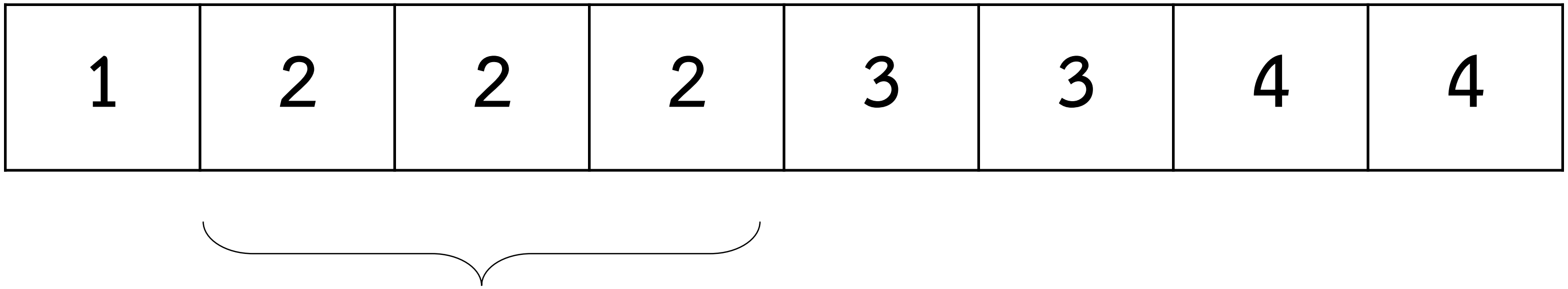
  - One example where the harmonic mean is useful is when examining the speed for a number of fixed-distance trips. For example, if the speed for going from point A to B was 60 km/h, and the speed for returning from B to A was 40 km/h, then the harmonic mean speed is given by

$$\frac{2}{\frac{1}{60}+\frac{1}{40}} = 48$$

Ref: • https://en.wikipedia.org/wiki/Average

Mode

- The most frequently occurring number in a list is called the mode.



1	2	2	2	3	3	4	4

Mode = 2

Ref: • https://en.wikipedia.org/wiki/Average

Median

- The median is the middle number of the group when they are ranked in order.

1	2	2	2	3	3	4	4	5

Median = 3

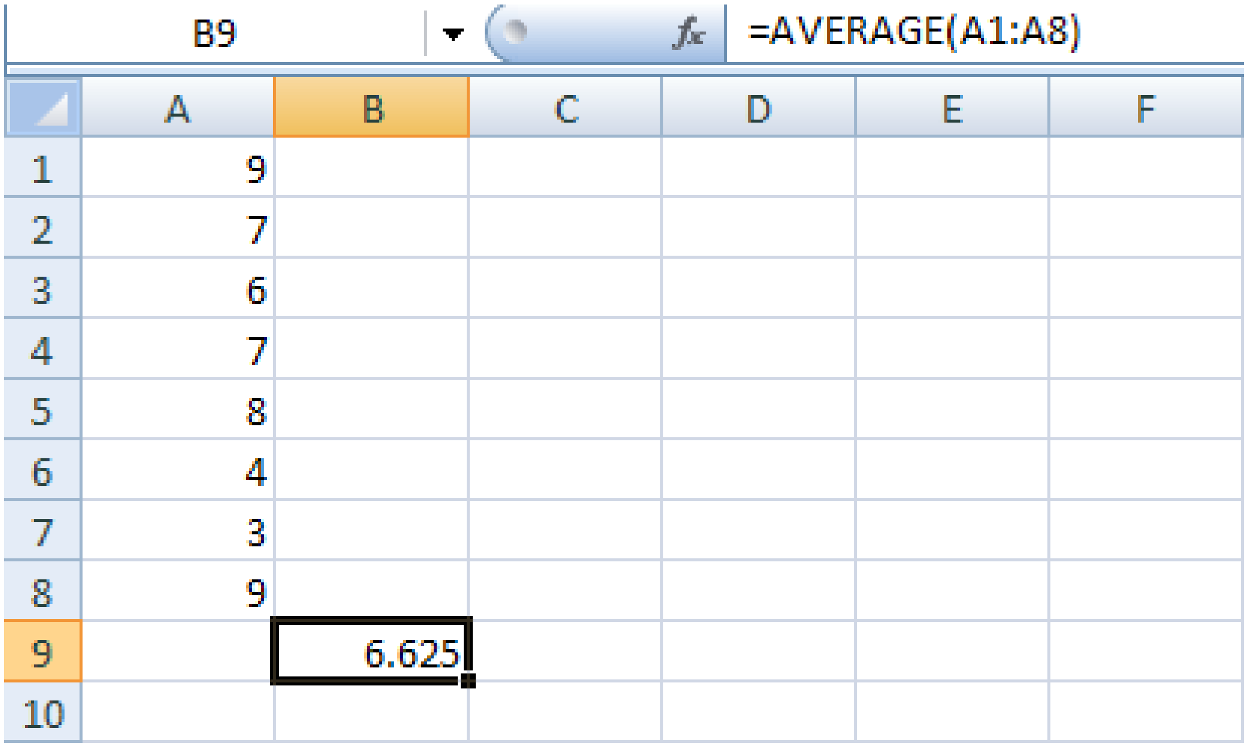
1	2	2	2	3	3	4	4

Median = (2+3)/2 = 2.5

Ref: • https://en.wikipedia.org/wiki/Average

Excel

- Mean



B9 =AVERAGE(A1:A8)

	A	B	C	D	E	F
1	9					
2	7					
3	6					
4	7					
5	8					
6	4					
7	3					
8	9					
9		6.625				
10						

Ref: • (image) https://www.homeandlearn.co.uk/excel2007/excel2007s5p1.html

Measure of Dispersion

Mean?

40	40	60	60

50	50	50	50

0	20	80	100

$$\left.\begin{array}{l}\bar{x} = 50\\[1em]\bar{x} = 50\\[1em]\bar{x} = 50\end{array}\right\}$$

is mean ok?

Range

$$R = Max - Min$$

40	40	60	60

$$R = 20$$

50	50	50	50

$$R = 0$$

0	20	80	100

$$R = 100$$

0	1	2	100

$$R = 100$$

Variance

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

40	40	60	60

$$S = 100$$

50	50	50	50

$$S = 0$$

0	20	80	100

$$S = 1700$$

Standard Deviation

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

40	40	60	60

$s = 10$

50	50	50	50

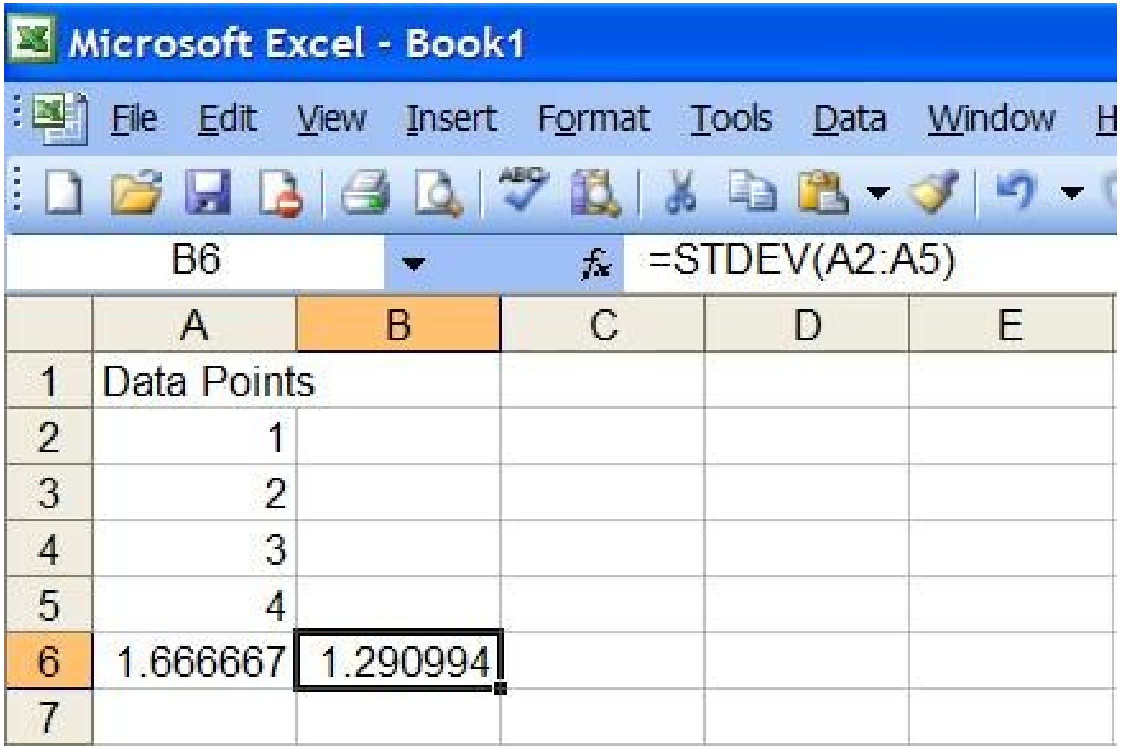
$s = 0$

0	20	80	100

$s = 41.23$

Excel

- Standard Deviation

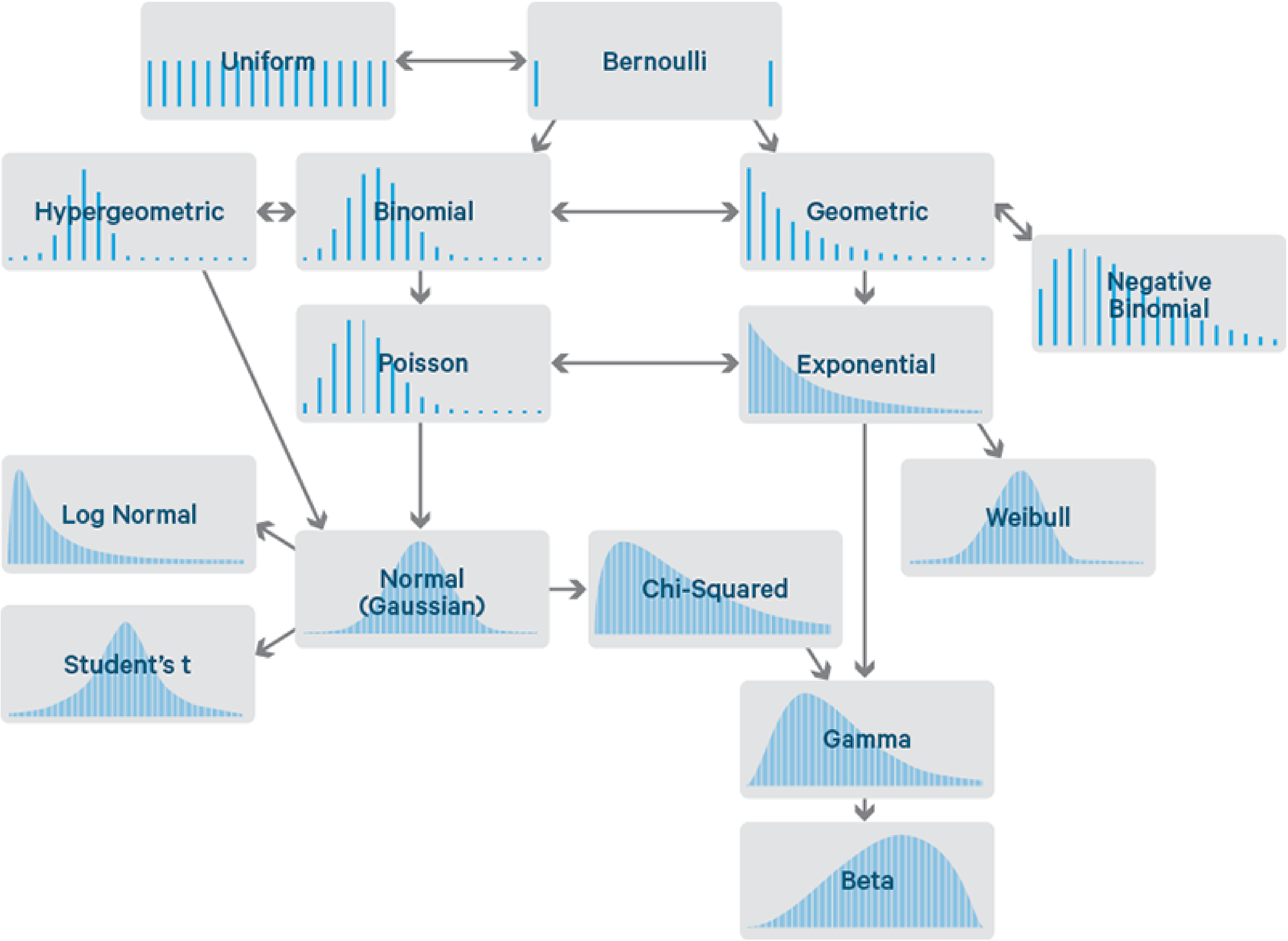


B6 f_x =STDEV(A2:A5)

	A	B	C	D	E
1	Data Points				
2	1				
3	2				
4	3				
5	4				
6	1.666667	1.290994			
7					

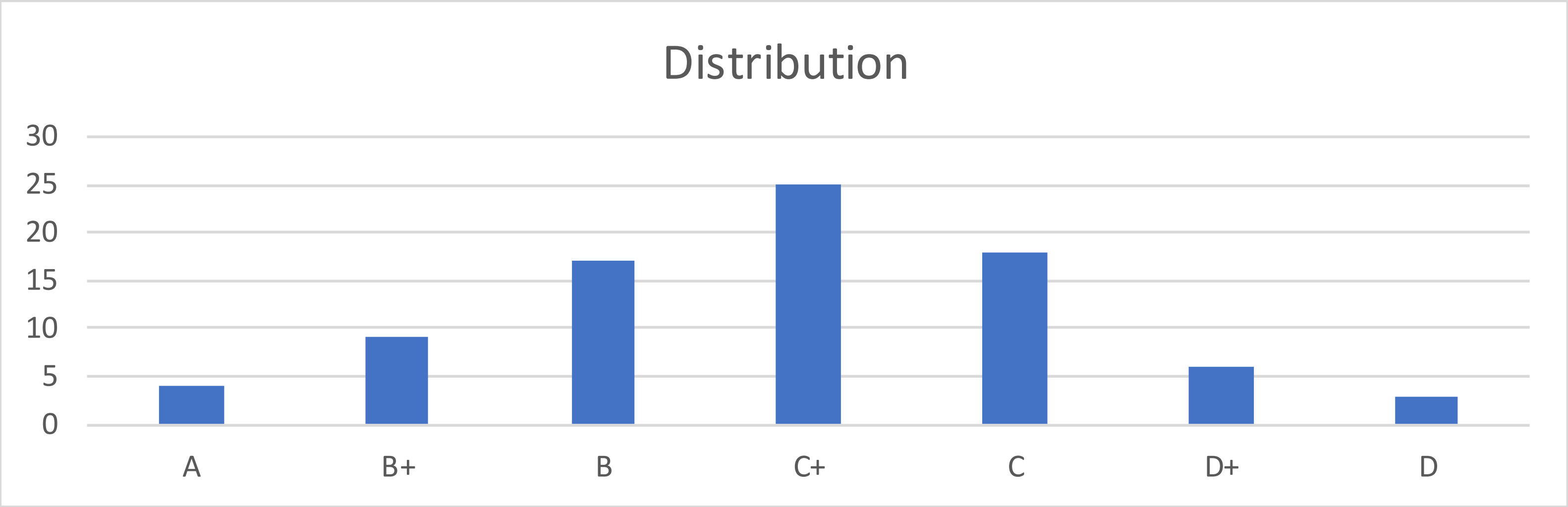
Ref: • (image) http://sciencefair.math.iit.edu/analysis/statistics/

Distribution



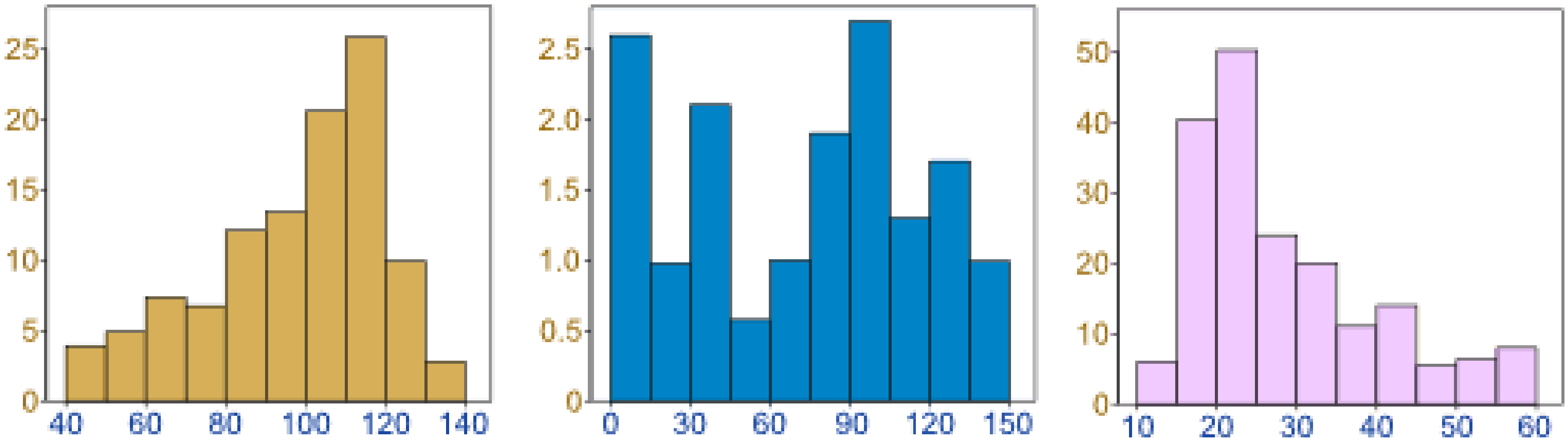
Ref: • (image) http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

ณ วิชาหนึ่ง



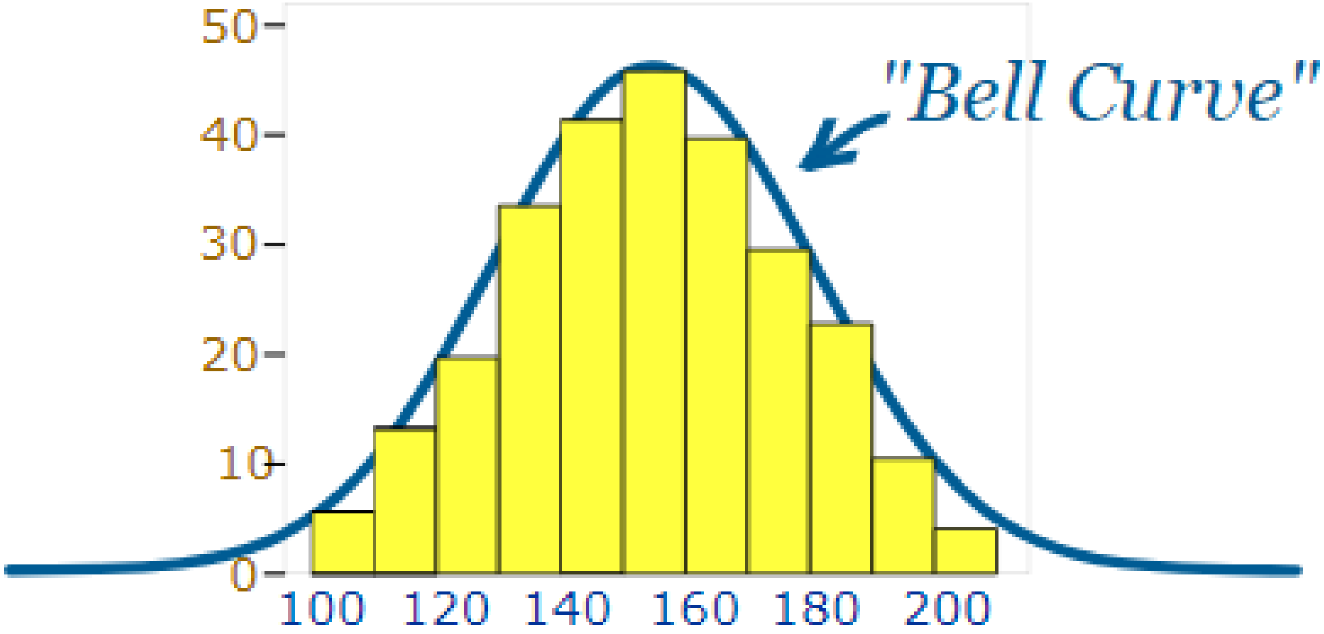
Grade	A	B+	B	C+	C	D+	D
Number	4	9	17	25	18	6	3

Distribution



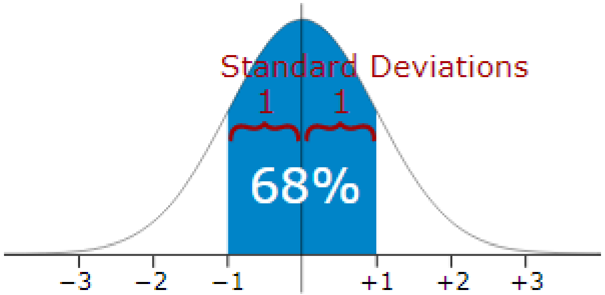
Ref: • (image) https://www.mathsisfun.com/data/standard-normal-distribution.html

Normal Distribution

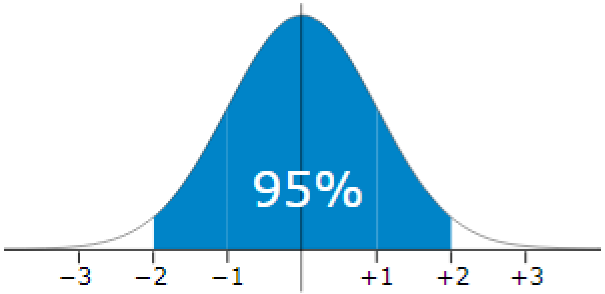


Ref: • (image) https://www.mathsisfun.com/data/standard-normal-distribution.html

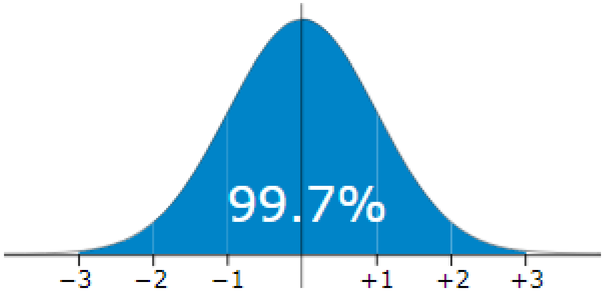
Normal Distribution



68% of values are within
1 standard deviation of the mean



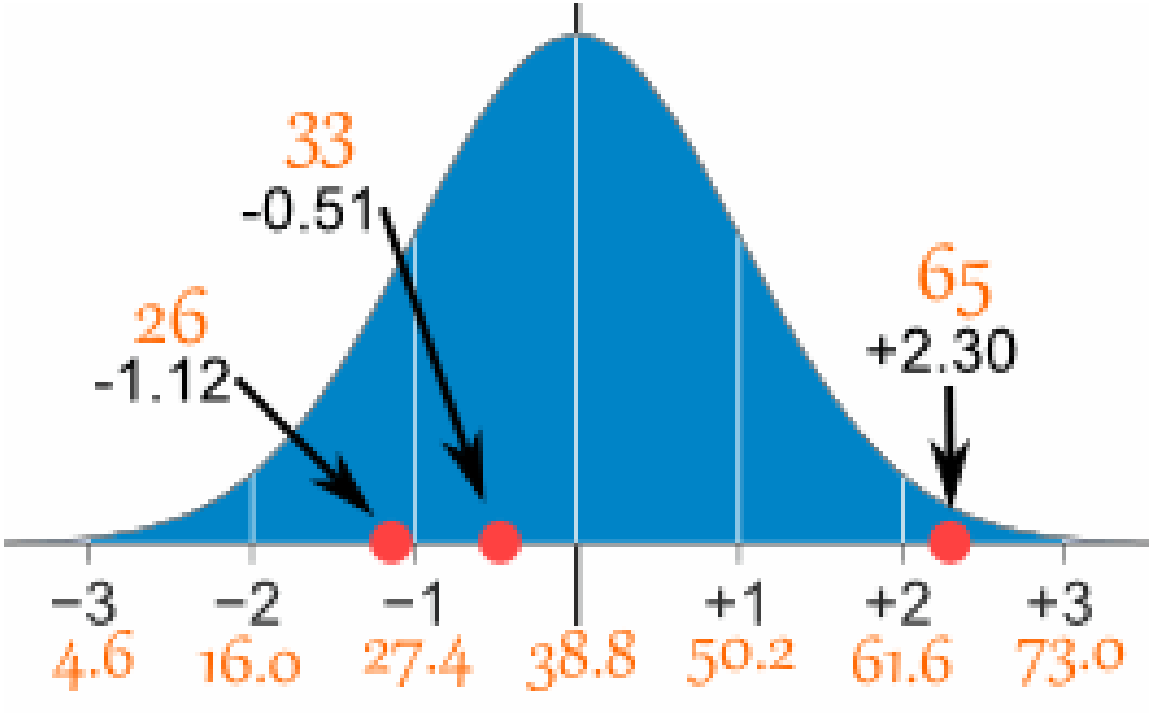
95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

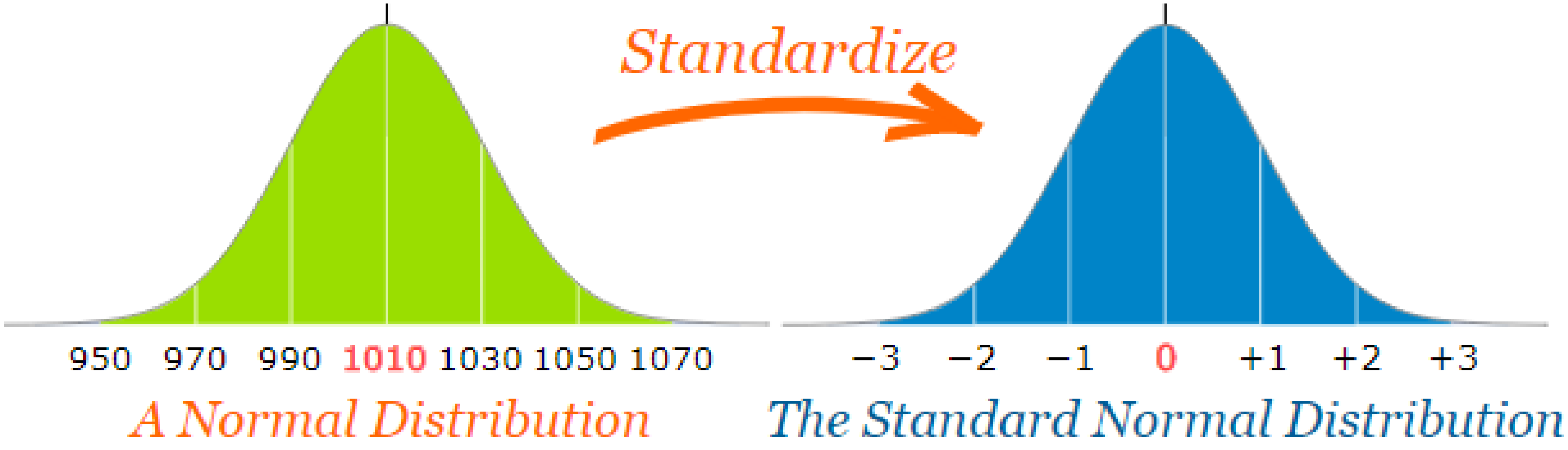
Ref: • (image) https://www.mathsisfun.com/data/standard-normal-distribution.html

Normal Distribution



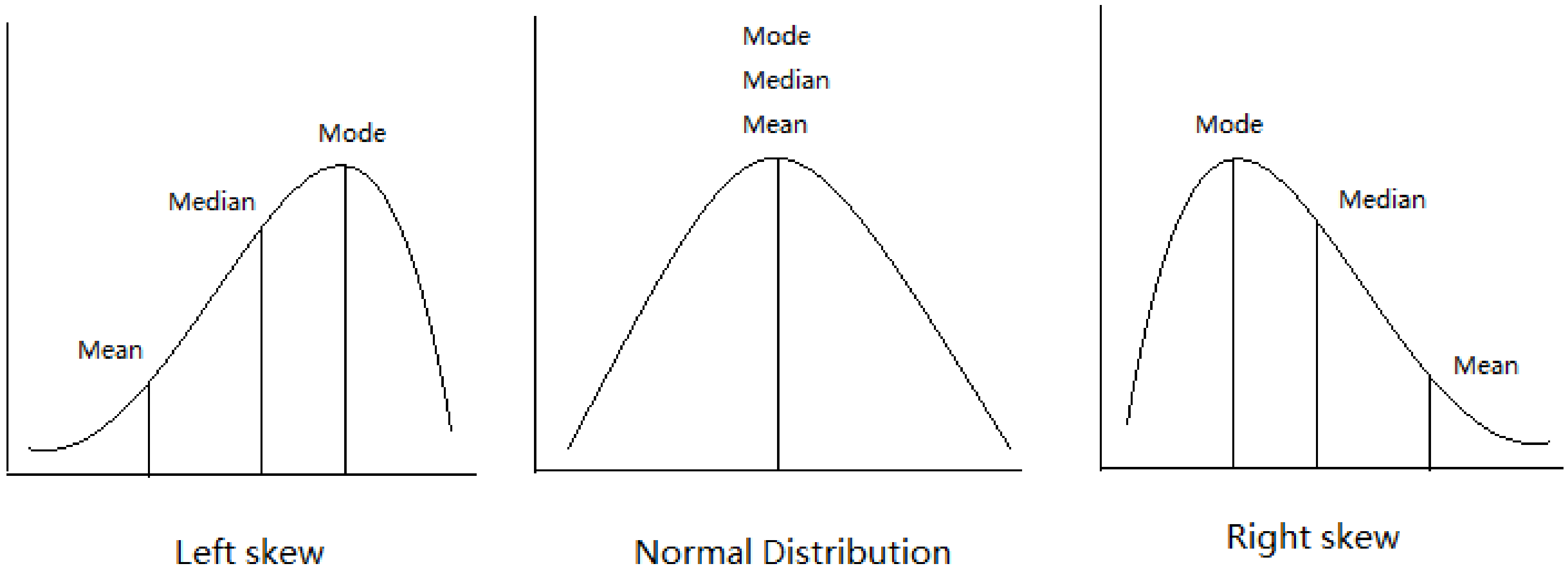
Ref: • (image) https://www.mathsisfun.com/data/standard-normal-distribution.html

Standardize



Ref: • (image) https://www.mathsisfun.com/data/standard-normal-distribution.html

Mean, Median, Mode



Left skew

Normal Distribution

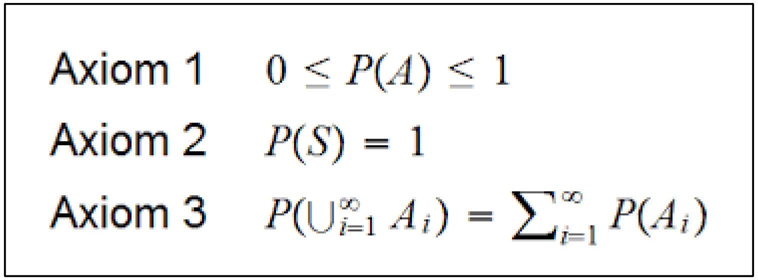
Right skew

Ref: • (image) http://access-excel.tips/excel-central-tendency-mean-mode-median/

Probability

Probability

- **Sample space** – The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by **S**.
- **Event** – Any subset ***E*** of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in ***E***, then we say that ***E*** has occurred.
- **Axioms of probability** – For each event ***E,*** we denote ***P(E)*** as the probability of event ***E*** occurring.
  By noting ***E1,…,En*** mutually exclusive events, we have the 3 following axioms



Axiom 1 $0 \leq P(A) \leq 1$

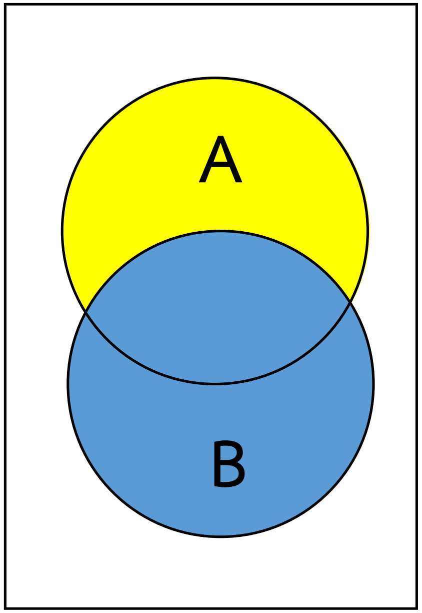
Axiom 2 $P(S) = 1$

Axiom 3 $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Ref: • (content) URL
• (image) URL

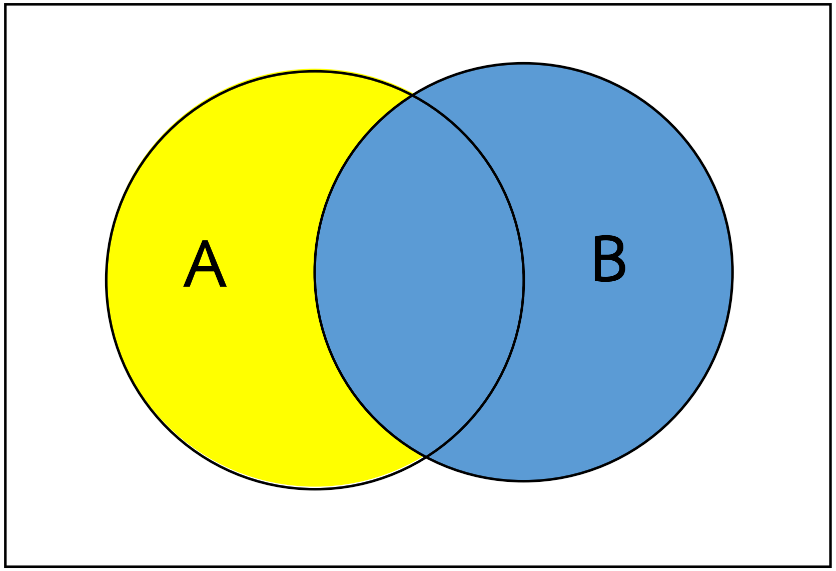
Rules of Probability

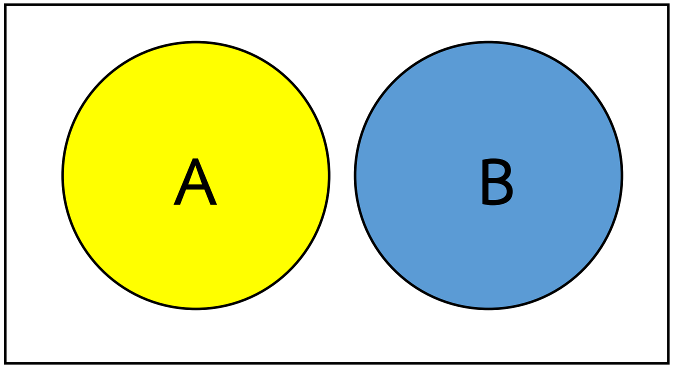
- Rule of Addition
  - The probability that Event A or Event B occurs is equal to the probability that Event A occurs plus the probability that Event B occurs minus the probability that both Events A and B occur.
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Rule of Subtraction.**
  - The probability that event A will occur is equal to 1 minus the probability that event A will not occur.
  - $P(A) = 1 - P(A')$
- **Rule of Multiplication**
  - The probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred.
  - $P(A \cap B) = P(A)\,P(B|A)$



Ref: • (content) https://stattrek.com/probability/probability-rules.aspx

Independent Event





Mutually Exclusive Event

Independent Event

$$P\,(A \cap B) = P\,(A\,)P(B)$$

Ref:
- (content) URL
- (image) URL

Conditional Probability

trying to find

know

Probability of event A occured and event B occured

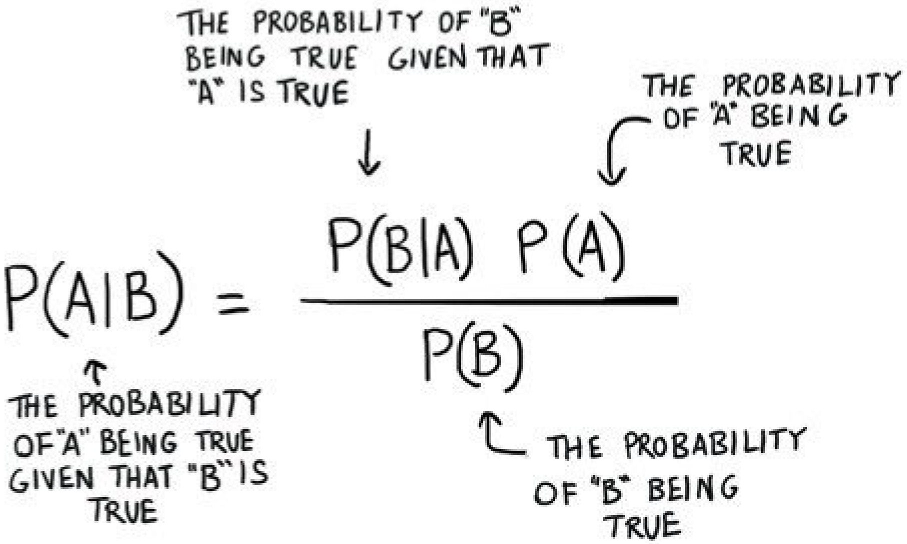
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

read "probability of A given B"

Probability of event B

Ref:
- (content) URL
- (image) URL

Bayes


THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE
THE PROBABILITY OF "A" BEING TRUE
P(A|B) = P(B|A) P(A) / P(B)
THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE
THE PROBABILITY OF "B" BEING TRUE

- Which tells us:
  - how often A happens given that B happens, written **P(A|B)**,
- When we know:
  - how often B happens given that A happens, written P(B|A)
  - and how likely A is on its own, written **P(A)**
  - and how likely B is on its own, written **P(B)**

Ref:
- (content) https://www.mathsisfun.com/data/bayes-theorem.html
- (image) https://pathmind.com/wiki/bayes-theorem-naive-bayes

Example 1

- How often there is fire when we can see smoke
  If dangerous fires are rare (1%)
  but smoke is fairly common (10%) due to barbecues,
  and 90% of dangerous fires make smoke

Ref: • (content) https://www.mathsisfun.com/data/bayes-theorem.html

Example 1

- How often there is fire when we can see smoke
  If dangerous fires are rare (1%)
  but smoke is fairly common (10%) due to barbecues,
  and 90% of dangerous fires make smoke
- ANS
  - P(Fire) means how often there is fire
  - P(Smoke) means how often we see smoke
  - P(Fire|Smoke) means how often there is fire when we can see smoke
  - P(Smoke|Fire) means how often we can see smoke when there is fire

$$P(Fire|Smoke) = \frac{P(Fire)P(Smoke|Fire)}{P(Smoke)} = \frac{0.01 \times 0.90}{0.10} = 0.09$$

Ref: • (content) https://www.mathsisfun.com/data/bayes-theorem.html

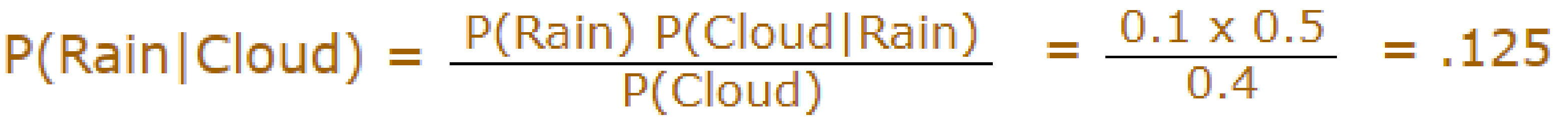
Example 2

- Picnic Day
  - You are planning a picnic today, but the morning is cloudy
    - Oh no! 50% of all rainy days start off cloudy!
    - But cloudy mornings are common (about 40% of days start cloudy)
    - And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)
  - What is the chance of rain during the day?

Ref: • (content) https://www.mathsisfun.com/data/bayes-theorem.html

Example 2

- Picnic Day
  - You are planning a picnic today, but the morning is cloudy
    - Oh no! 50% of all rainy days start off cloudy!
    - But cloudy mornings are common (about 40% of days start cloudy)
    - And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)
  - What is the chance of rain during the day?
- ANS
  - We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.
  - The chance of Rain given Cloud is written **P(Rain|Cloud)**
  - **P(Rain)** is Probability of Rain = 10%
  - **P(Cloud|Rain)** is Probability of Cloud, given that Rain happens = 50%
  - **P(Cloud)** is Probability of Cloud = 40%



$$P(\text{Rain}|\text{Cloud}) = \frac{P(\text{Rain})\,P(\text{Cloud}|\text{Rain})}{P(\text{Cloud})} = \frac{0.1 \times 0.5}{0.4} = .125$$

Ref: • (content) https://www.mathsisfun.com/data/bayes-theorem.html

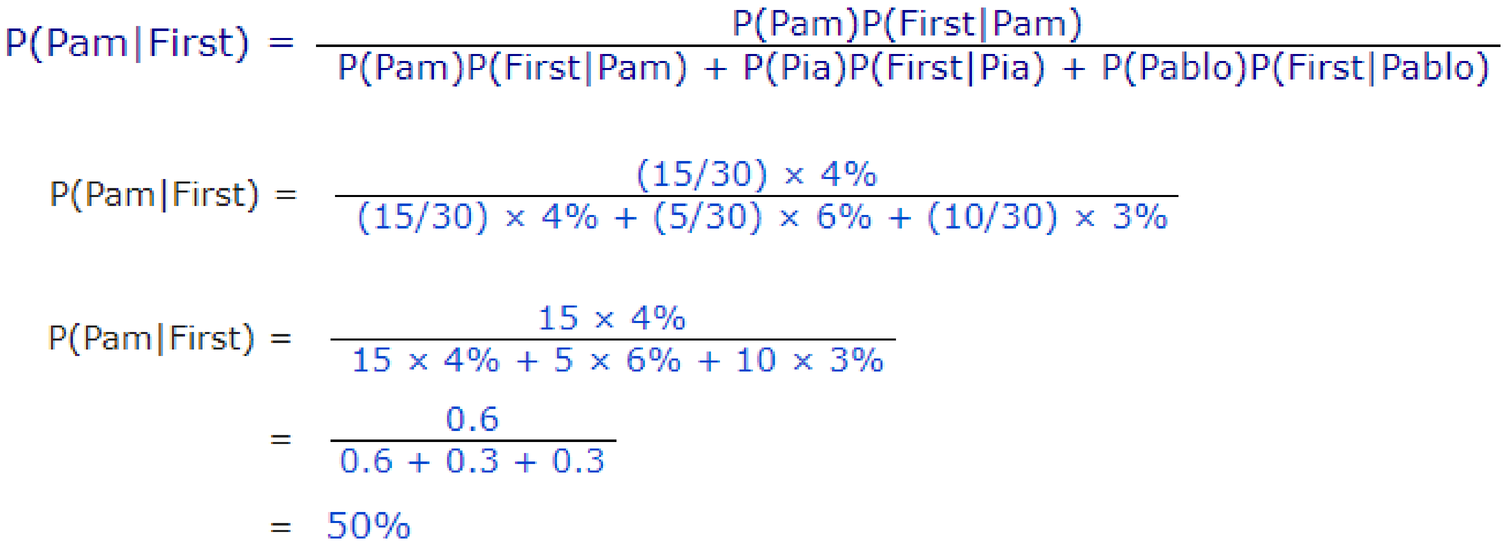
Example 3

- The Art Competition has entries from three painters: Pam, Pia and Pablo
  - Pam put in 15 paintings, 4% of her works have won First Prize.
  - Pia put in 5 paintings, 6% of her works have won First Prize.
  - Pablo put in 10 paintings, 3% of his works have won First Prize.
- What is the chance that Pam will win First Prize?

Ref: • (content) https://www.mathsisfun.com/data/bayes-theorem.html

Example 3

- The Art Competition has entries from three painters: Pam, Pia and Pablo
  - Pam put in 15 paintings, 4% of her works have won First Prize.
  - Pia put in 5 paintings, 6% of her works have won First Prize.
  - Pablo put in 10 paintings, 3% of his works have won First Prize.
- What is the chance that Pam will win First Prize?



$$P(\text{Pam}|\text{First}) = \frac{P(\text{Pam})P(\text{First}|\text{Pam})}{P(\text{Pam})P(\text{First}|\text{Pam}) + P(\text{Pia})P(\text{First}|\text{Pia}) + P(\text{Pablo})P(\text{First}|\text{Pablo})}$$

$$P(\text{Pam}|\text{First}) = \frac{(15/30) \times 4\%}{(15/30) \times 4\% + (5/30) \times 6\% + (10/30) \times 3\%}$$

$$P(\text{Pam}|\text{First}) = \frac{15 \times 4\%}{15 \times 4\% + 5 \times 6\% + 10 \times 3\%}$$

$$= \frac{0.6}{0.6 + 0.3 + 0.3}$$

$$= 50\%$$

Ref: • (content) https://www.mathsisfun.com/data/bayes-theorem.html

Example 4

- Three machines, M1, M2 y M3, produce 45%, 30%, and 25%, respectively, of the total parts produced in a factory. The percentages of defective production of these machines are 3%, 4% y 5%, respectively.
  - a) If we choose a part randomly, calculate the probability that it is defective.
  - b) Suppose now that we choose a part randomly and it is defective. Calculate the probability that it was produced by M2.

Ref: • (content) http://www.mathspadilla.com/macsII/Unit9-Probability/bayes_theorem.html

Example 4

- Three machines, M1, M2 y M3, produce 45%, 30%, and 25%, respectively, of the total parts produced in a factory. The percentages of defective production of these machines are 3%, 4% y 5%, respectively.
  - a) If we choose a part randomly, calculate the probability that it is defective.

$$\left[P(D)=P(M_1)\cdot P\left(D/M_1\right)+P(M_2)\cdot P\left(D/M_2\right)+P(M_3)\cdot P\left(D/M_3\right)=0.45\cdot 0.03+0.3\cdot 0.04+0.25\cdot 0.05=0.038\right]$$

  - b) Suppose now that we choose a part randomly and it is defective. Calculate the probability that it was produced by M2.

$$\left[P(M_2/D)=\frac{P(M_2)\cdot P(D/M_2)}{P(M_1)\cdot P(D/M_1)+P(M_2)\cdot P(D/M_2)+P(M_3)\cdot P(D/M_3)}=\frac{0.3\cdot 0.04}{0.038}=0.316\right]$$

Ref: • (content) http://www.mathspadilla.com/macsII/Unit9-Probability/bayes_theorem.html

Other Questions

1. Three companies A, B and C supply 25%, 35% and 40% of the notebooks to a school. Past experience shows that 5%, 4% and 2% of the notebooks produced by these companies are defective. If a notebook was found to be defective, what is the probability that the notebook was supplied by A? *(answer = 25/69)*

2. An urn B1 contains 2 white and 3 black chips and another urn B2 contains 3 white and 4 black chips. One urn is selected at random and a chip is drawn from it. If the chip drawn is found black, find the probability that the urn chosen was B1. *(answer = 21/41)*

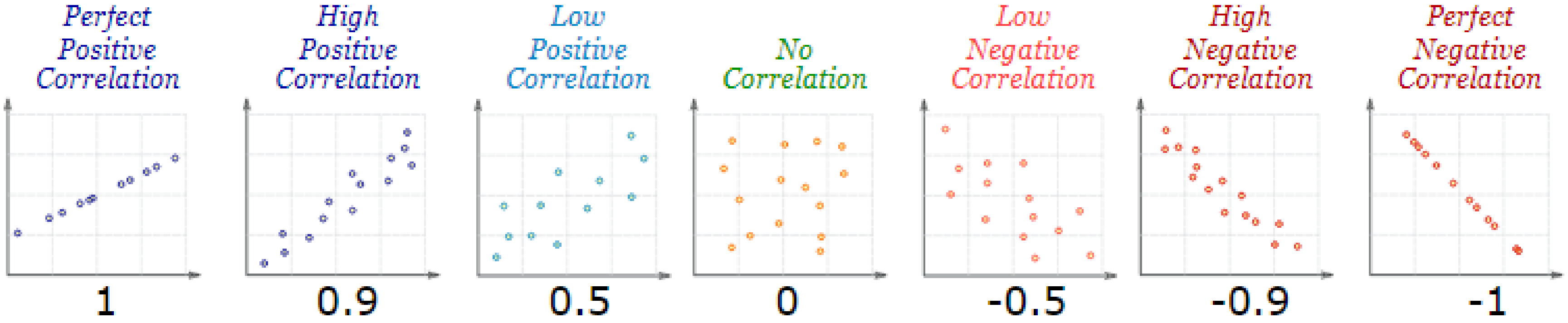
3. At a certain university, 4% of men are over 6 feet tall and 1% of women are over 6 feet tall. The total student population is divided in the ratio 3:2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman? *(answer = 3/11)*

Ref: • (content) https://www.sanfoundry.com/probability-statistics-questions-answers-bayes-theorem/

Correlation

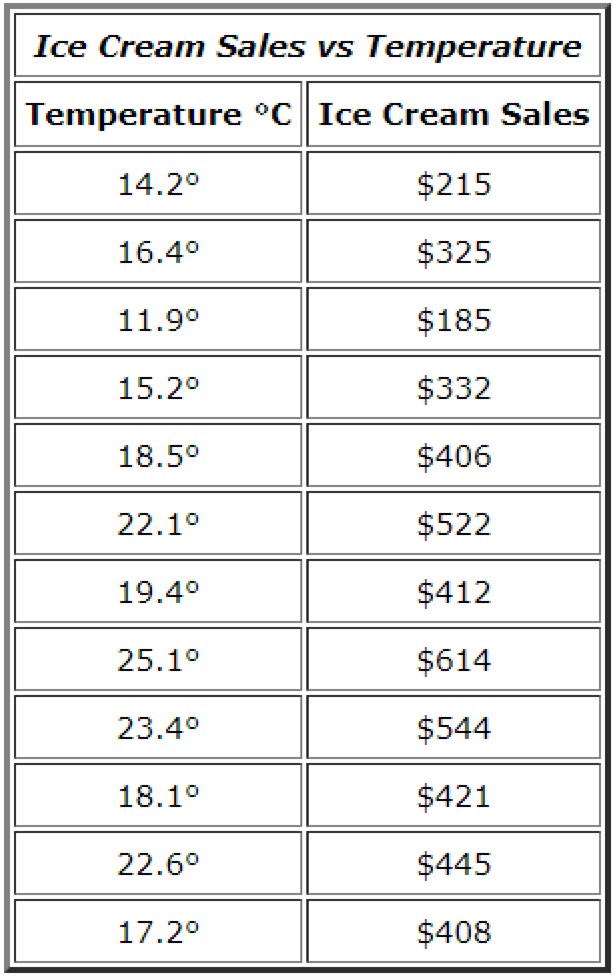
Correlation

- When two sets of data are strongly linked together we say they have a High Correlation.
  - Correlation is Positive when the values increase together, and
  - Correlation is Negative when one value decreases as the other increases

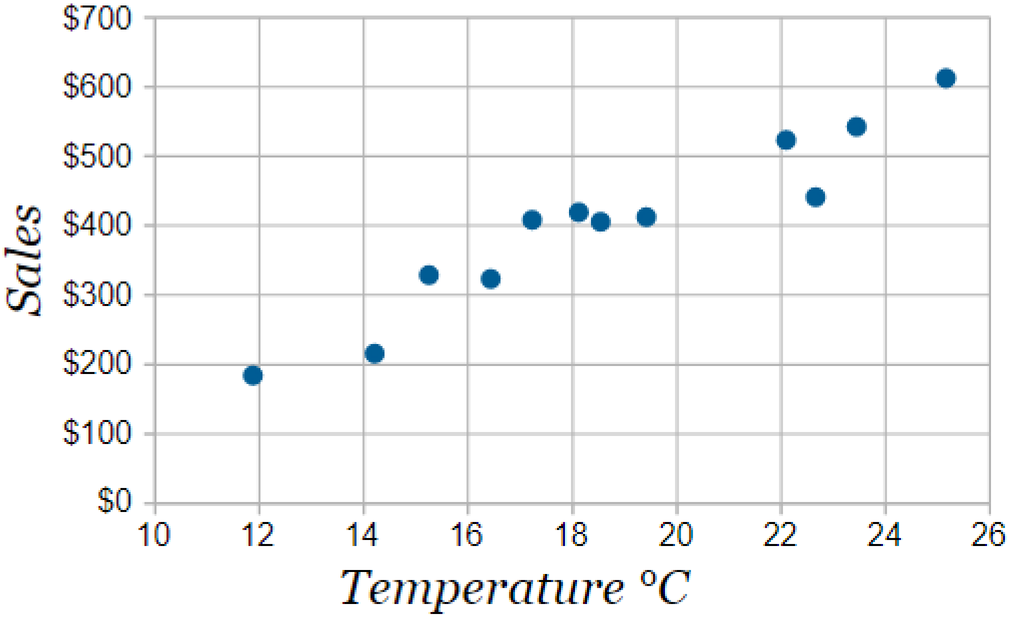


Ref: • https://www.mathsisfun.com/data/correlation.html

Example: Ice-cream Sales & Temperature



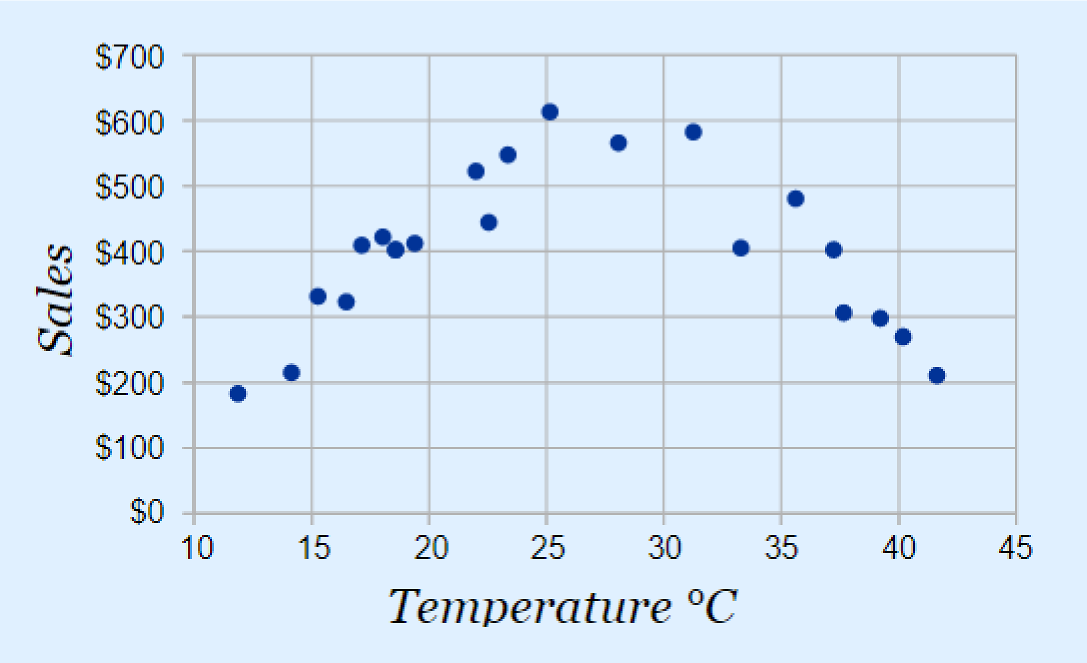
Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



- he correlation is 0.9575

Ref: • https://www.mathsisfun.com/data/correlation.html

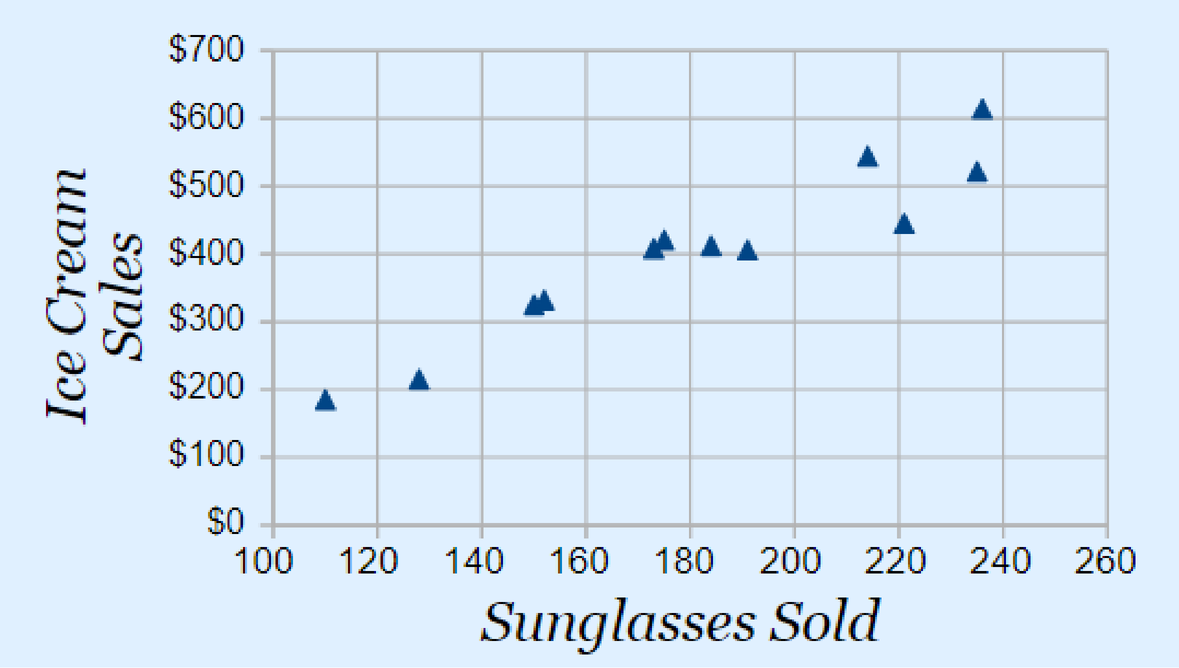
Correlation Is Not Good at Curves



- The correlation value is now 0: "No Correlation" ... !
- Our Ice Cream Example: there has been a heat wave!
- It gets so hot that people aren't going near the shop, and sales start dropping.

Ref: • https://www.mathsisfun.com/data/correlation.html

Correlation Is Not Causation



- The correlation between Sunglasses and Ice Cream sales is high
- "Correlation Is Not Causation" ... which says that a correlation does not mean that one thing causes the other (there could be other reasons the data has a good correlation).

Ref: • https://www.mathsisfun.com/data/correlation.html

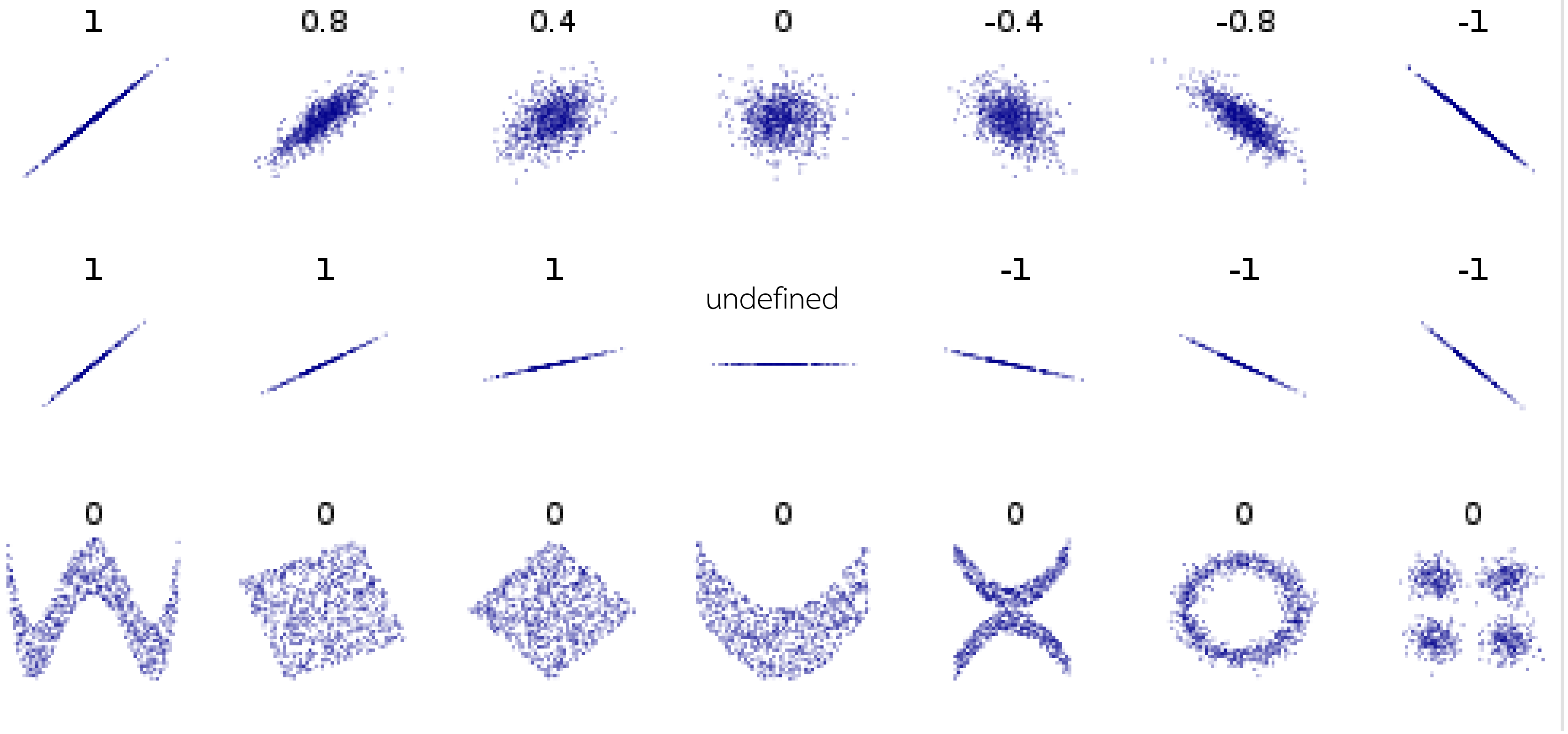
Pearson Correlation

- The most familiar measure of dependence between two quantities is the Pearson product-moment correlation coefficient, or "Pearson's correlation coefficient", commonly called simply "the correlation coefficient". It is obtained by dividing the covariance of the two variables by the product of their standard deviations. Karl Pearson developed the coefficient from a similar but slightly different idea by Francis Galton.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Ref: • https://en.wikipedia.org/wiki/Correlation_and_dependence

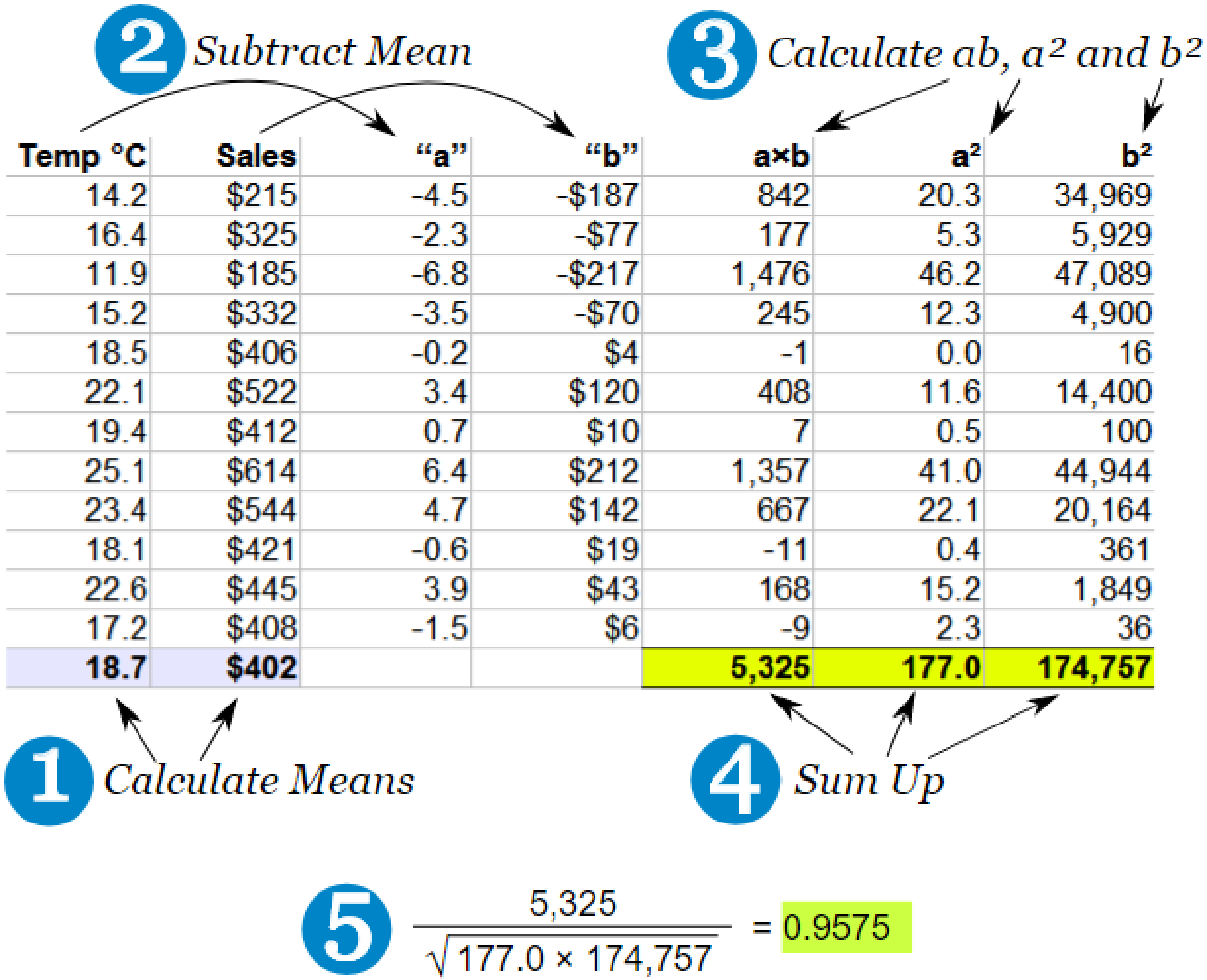
Pearson Correlation



Ref: • https://en.wikipedia.org/wiki/Correlation_and_dependence

Excel

- Pearson Correlation



2 *Subtract Mean*

3 *Calculate ab, a² and b²*

Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

1 *Calculate Means*

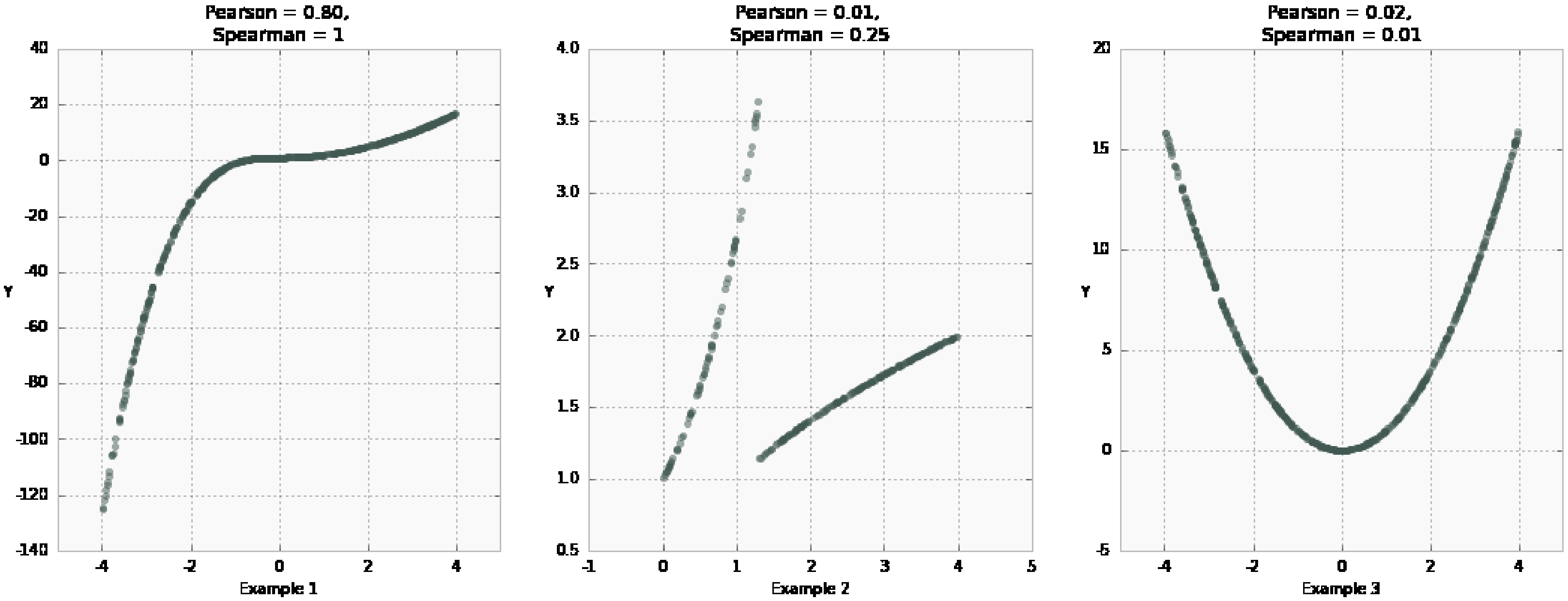
4 *Sum Up*

5 $$\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$$

Ref: • https://www.mathsisfun.com/data/correlation.html

Other Correlation Indexes

- Rank correlation coefficients, such as Spearman's rank correlation coefficient and Kendall's rank correlation coefficient (τ) measure the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship.



Ref:
- (content) https://en.wikipedia.org/wiki/Correlation_and_dependence
- (image) https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

By a small sample, we may judge of the whole piece.

Miguel de Cervantes

つづく