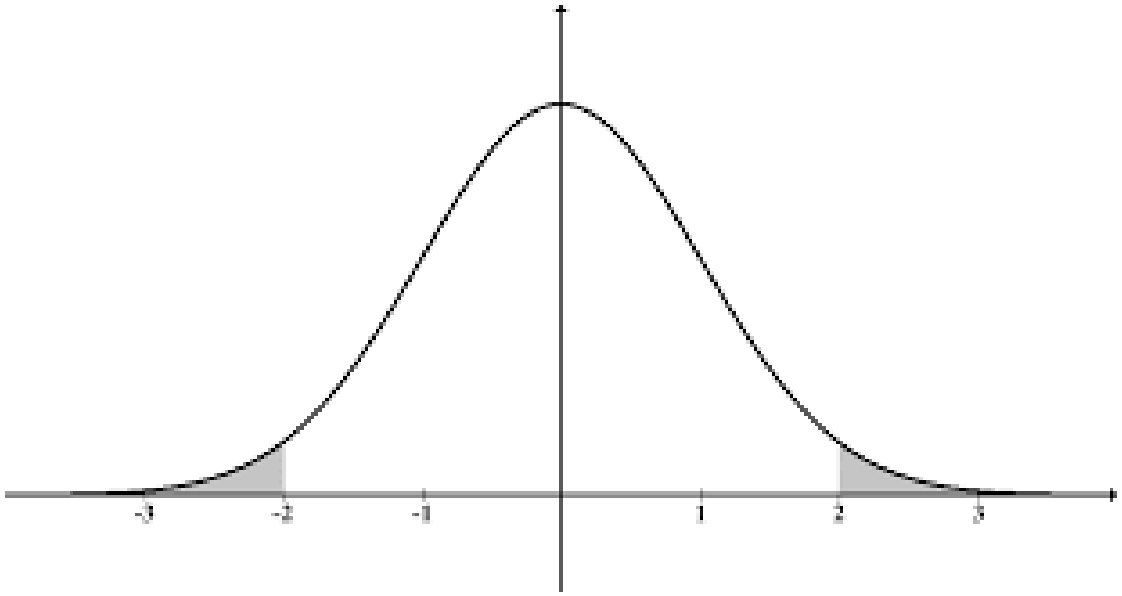
Statistics Fundamental Training v8



Trainer: Marcus Lee

Website:www.tertiarycourses.com.sg
Email: enquiry@tertiaryinfotech.com

About the Trainer

Marcus Lee has a degree in Computer Science and a background in Statistics from the University of Otago. Before returning to Singapore, he analyzed vacation data provided by the New Zealand Board of Tourism to determine the favorite activities of Australian, Japanese, and German tourists in New Zealand. In addition to a vast number of other demographics statistics, he has been able to provide significant advice to the board on how to promote tourism in New Zealand. His core specialization skills are Java, R, Statistical Analysis, Machine Learning, NumPy, Scikit, and Network Management. He has also a fair amount of experience in C, C++, and Python

Let's Know Each Other...

Say a bit about yourself

- Name
- What Industry you are from?
- Do you have any prior knowledge in statistics or data analysis?
- Why do you want to learn statistics?

Ground Rules

- Set your mobile phone to silent mode
- Actively participate in the class. No question is stupid.
- Respect each other view
- Exit the class silently if you need to step out for phone call, toilet break

Ground Rules for Virtual Training

- Upon entering, mute your mic and turn on the video. Use a headset if you can
- Use the 'raise hand' function to indicate when you want to speak
- Participant actively. Feel free to ask questions on the chat whenever.
- Facilitators can use breakout rooms for private sessions.

Guidelines for Facilitators

1. Once all the participants are in and introduce themselves
2. Go to gallery mode, take a snapshot of the class photo - makes sure capture the date and time
3. Start the video recording (only for WSQ courses)
4. Continue the class
5. Before the class end on that day, take another snapshot of the class photo - makes sure capture the date and time
6. For NRIC verification, facilitator to create breakout room for individual participant to check (only for WSQ courses)
7. Before the assessment start, take another snapshot of the class photo - makes sure capture the date and time (only for WSQ courses)
8. For Oral Questioning assessment, facilitator to create breakout room for individual participant to OQ (only for WSQ courses)
9. End the video recording and upload to cloud (only for WSQ courses)
10. Assessor to send all the assessment records, assessment plan and photo and video to the staff (only for WSQ courses).

Prerequisite

This is a beginner course. No prerequisite is assumed.

Exercise Files (*)

Please go to the following link and download a copy of the exercise files:

https://github.com/makasulee0/statsData

Agenda

Topic 1 Introduction to Statistics

- Why Statistics Matter
- Categorical and Quantitative Data
- Descriptive Statistics: Mean and Standard Deviation
- Probability and Conditional Probability
- Bayes Theorem
- Discrete Probability Distributions
- Continuous Probability Distributions
- Software for Statistical Analysis

Topic 2 Sampling

- Sampling Consideration
- Central Limit Theorem
- Sampling Distribution of the Mean
- Standard Errors for Proportion and Mean
- Confidence Interval
- T-Statistics vs Z-Statistics
- T-Score Table and Degree of Freedom
- Calculating Confidence Interval of T-Score

Agenda

Topic 3 Hypothesis Testing

- Overview of Hypothesis Testing
- Steps for Performing a Hypothesis Testing
- P-Value and Significance Level
- Types of Hypothesis Testing
- One Tailed vs Two Tailed Hypothesis Testing
- Type 1 and Type 2 Errors

Topic 4 Chi-Square Testing

- Overview of Chi-Square Hypothesis Testing
- Chi-Square Statistic and Distribution
- Goodness of Fit Test

Topic 5 ANOVA: Analysis of Variance

- What is Analysis of Variance
- F Statistics and Distribution
- One Way ANOVA
- Two Way ANOVA

Agenda

Topic 6 Regression

- What is Regression?
- Residues and Mean Square Error

Topic 7 Correlation Analysis

- Covariance and Covariance Matrix
- Correlation Coefficient and Correlation Matrix

Google Classroom

- Go to google classroom https://classroom.google.com
- Enter the class code below to join the class on the top right.
- If you cannot access the google classroom, please inform the trainer or staff.

fkr4vg6

Topic 1
Introduction to Statistics

What is Statistics?

Statistics is a discipline which is concerned with:

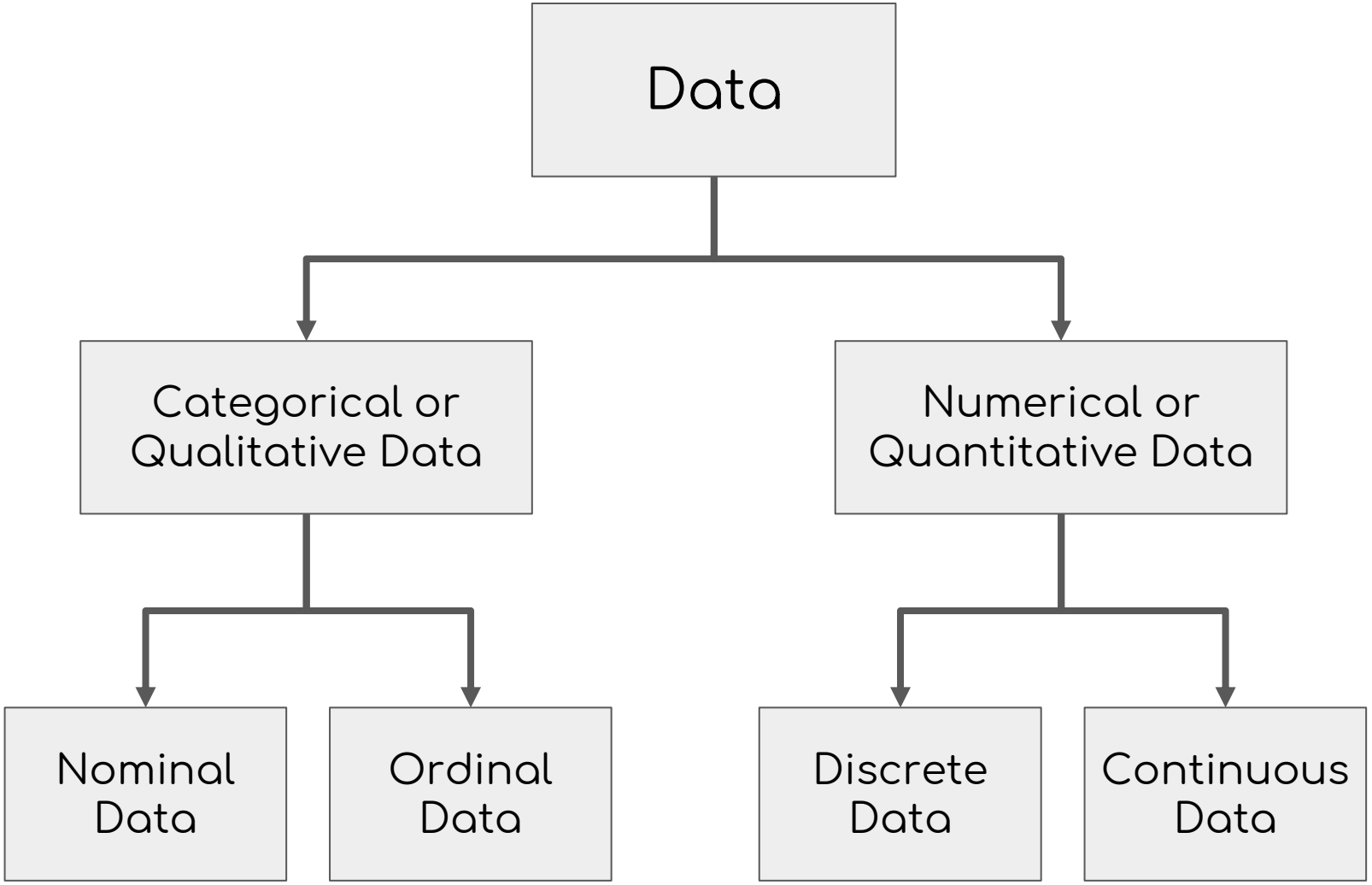
- designing experiments and other data collection,
- summarizing information to aid understanding,
- drawing conclusions from data, and
- estimating the present or predicting the future.

Statistical statements:
"I sleep for about eight hours per night on average"
"You are more likely to pass the exam if you start preparing earlier"

Why Statistics Matter?

- Environmental Study
  - Is Singapore getting hotter over last 10 years?
- Policy Study
  - Is more people using green transport such as Bicycles, Buses, Carpool, CNG Cars, Electric Cars, Electric Scooters
- Market Analysis
  - Is more people likely to take green transport if they've seen a recent TV advertisement for green transport?
- Public Transport
  - Is more people likely to commute by MRT if we have more MRT stations in the neighborhood?
- Health Care
  - Does air pollution from vehicles cause any health concern?
- Data Science
  - Statistics is fundamental for understanding Artificial Intelligence and Machine Learning.

Types of Data



Categorical and Quantitative Data

- **Categorical (Qualitative) Data** - each observation belongs to one of a set of categories. Examples:
  - Weather (Rainy /Sunny)
  - Air Pollutants (Ozone/Nitrogen Dioxide)
  - Gender (Male or Female)
  - Place of residence (HDB, Condo, ...)
  - Marital status (Married, Single,...)
- **Quantitative (Numerical) Data** - observations take numerical values. Examples:
  - Surface Air temperature
  - Weekly number of dengue cases
  - No. of days with rainfall in a month
  - Age
  - Number of cars
  - Weight

Nominal and Ordinal Data

- **Nominal Data** is defined as data that is used for naming or labelling variables, without any quantitative value. It is sometimes called "labels" data E.g.
  - Male/Female
  - Red/Green/Blue

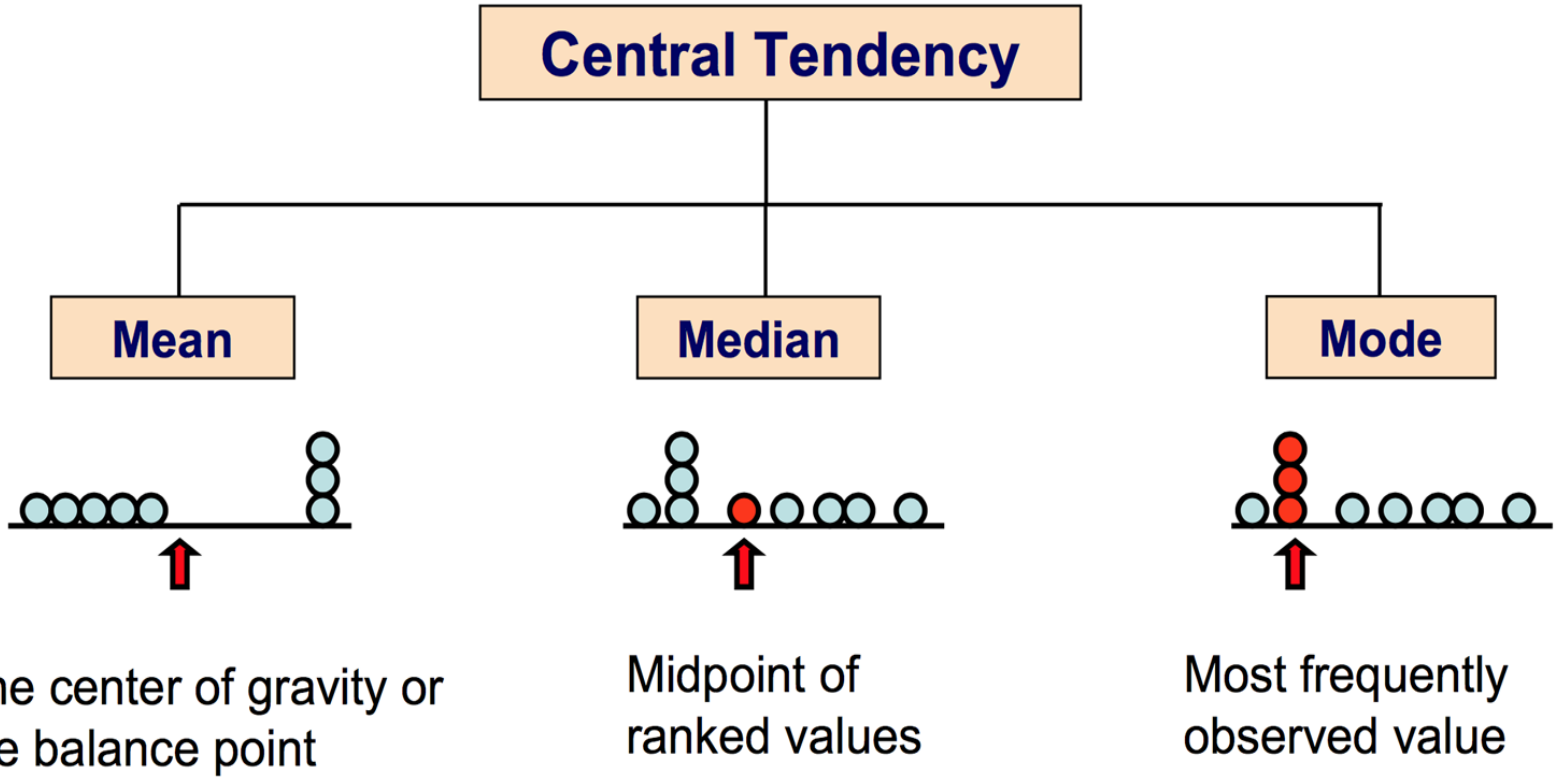
- **Ordinal Data** is a type of categorical data with an order. The variables in ordinal data are listed in an ordered manner.
  - Disagree/Neutral/Agree/Strongly Agree
  - Very Bad/Bad/Good/Very Good

Discrete and Continuous Data

- **Discrete Data is** a set of countable numbers such as 0, 1, 2, 3,.....Examples:
  - No. of days with rainfall in a month
  - Weekly no. of dengue cases
  - Number of children in a family
  - Number of foreign languages spoken

- **Continuous Data** are continuous numbers from an interval. Examples:
  - Surface Air temperature
  - Amount of rainfall in a month
  - Height
  - Weight

Measures of Central Tendency



- Mean - add up all the values and divide by how many there are
- Median - Arrange all the numbers from smallest to largest:
  - odd number of points: Median = middle value
  - even number of points: Median= mean of the middle two values

Measures of Central Tendency (*)

Consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

How do you describe the wages of staff at this factory?

Measures of Central Tendency: use a single value to describe a set of data by identifying the central position within that set of data.

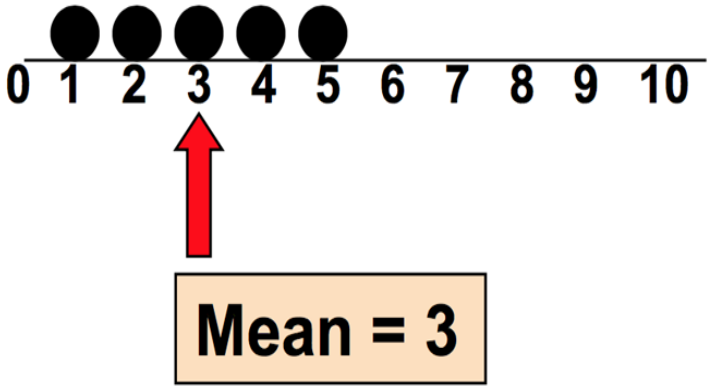
Mean (*)

The mean of a variable is often what people mean by the "average" ... add up all the values and divide by how many there are

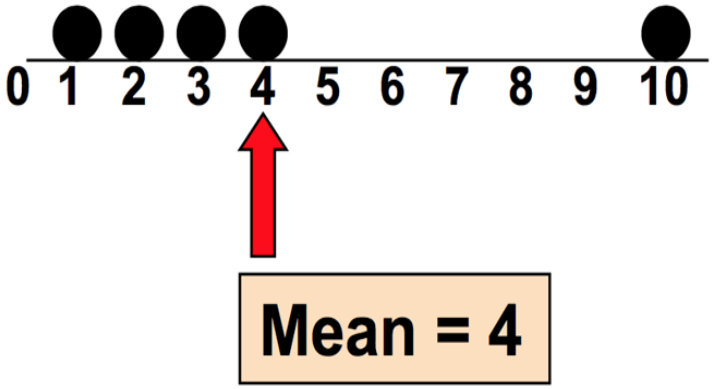
Example : Compute the mean of 6,1,5

Mean (*)

Affected by extreme values (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Median (*)

Arrange all the numbers from smallest to largest:

- odd number of points:

  Median = middle value

- even number of points:

  Median= mean of the middle two values

Median (*)

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= **4.5**

Median (*)

Example :

Compute the median of 6, 1, 11, 2, 11

(1, 2, **6**, 11, 11)

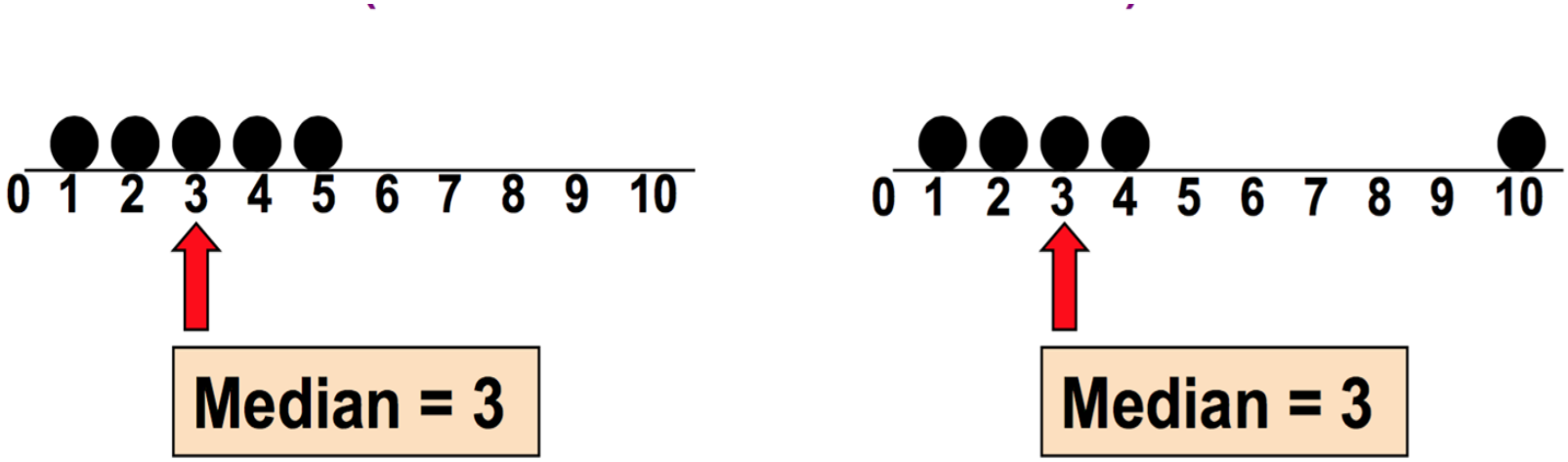
Example:

Compute the median of 6, 1, 11, 2

(1, **2**, **6**, 11)

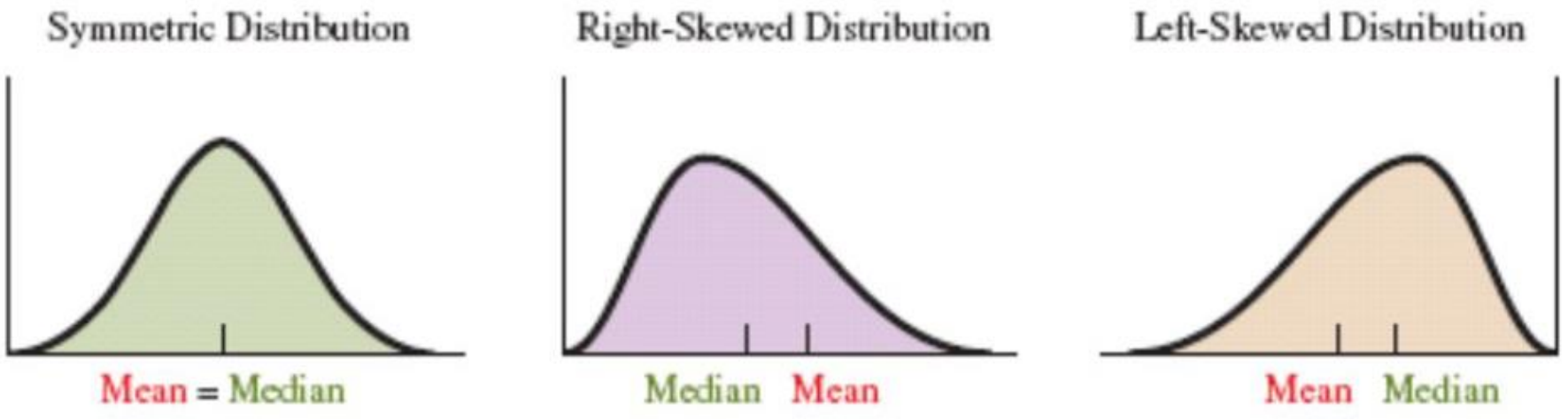
Median (*)

Not affected by extreme values



Mean vs Median

- Mean
  - Useful for roughly symmetric quantitative data
  - Sensitive to outlier data
- Median
  - Splits the data into halves
  - Useful for highly skewed quantitative data
  - Insensitive to outlier data



Exercise (*)

CO2 Pollution levels in 8 largest nations measured in metric tons per person:

2.3 1.1 19.7 9.8 1.8 1.2 0.7 0.2

mean=

median=

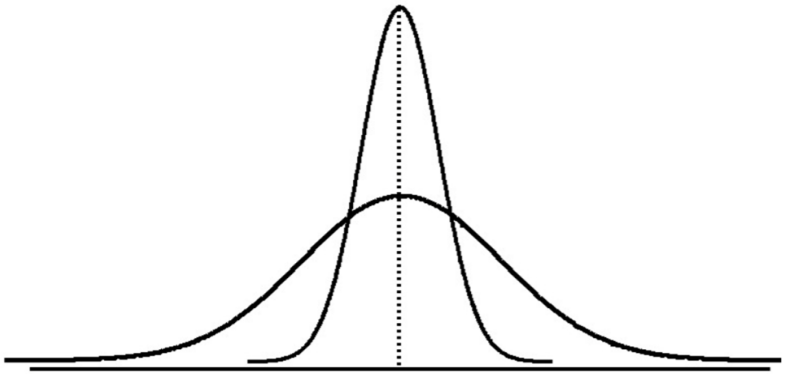
Module1a – CO2

Measures of Dispersion

- The measures of central tendency (mean, median) measure the differences between the "average" or "typical" values between two sets of data.

- The measures of dispersion measure the differences between how far "spread out" the data values are.

- Two commonly used measures for dispersion are: **range** and **standard deviation** (more commonly used).



Same center, different variation

Variance and Standard Deviation (*)

Sample Variance

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

Sample Standard Deviation

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

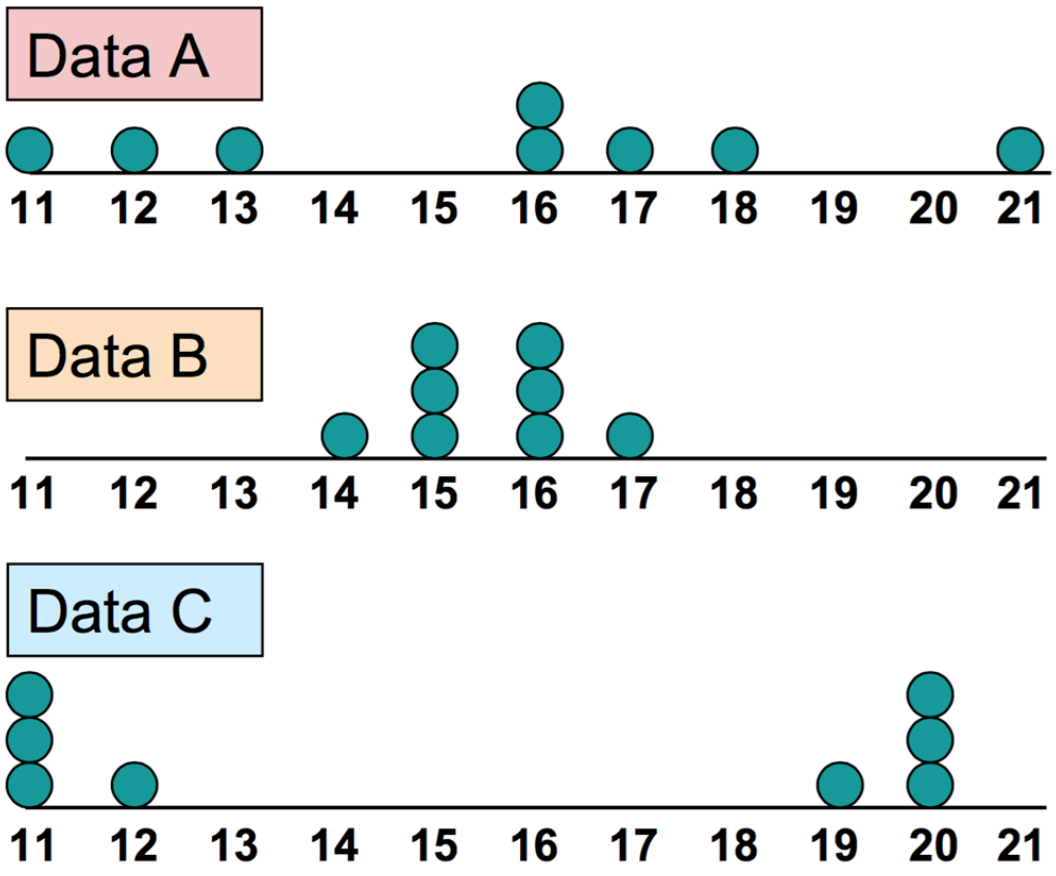
Standard Deviation

- The standard deviation measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

- Larger standard deviation (s) = Greater variability of the data

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

Compare Standard Deviations (*)



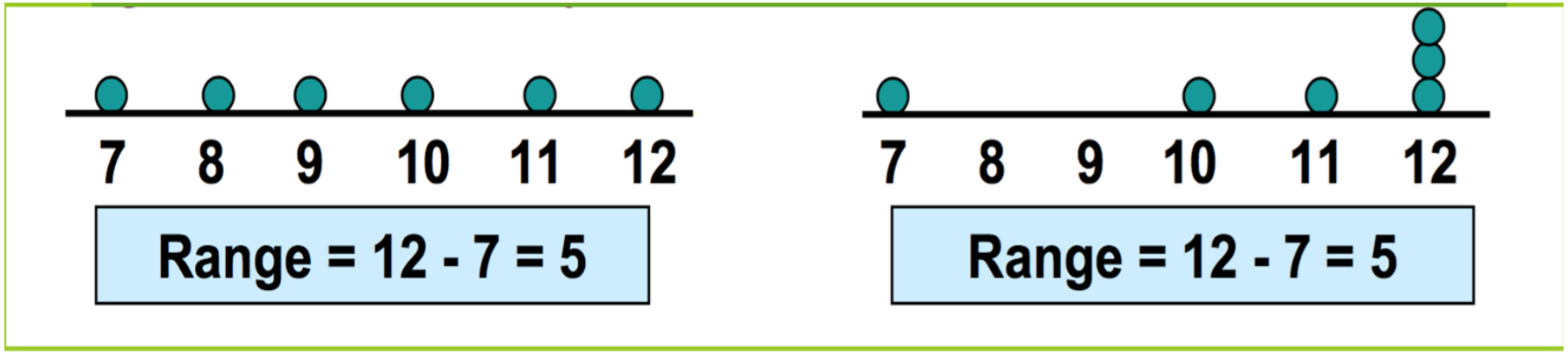
Mean = 15.5
s = 3.34

Mean = 15.5
s = 0.93

Mean = 15.5
s = 4.57

Range

- Range is the difference between the highest and lowest values.
- Since it uses only the extreme values, it is greatly affected by extreme values.
- Range ignores the way in which data are distributed.



Disadvantages of the Range (*)

Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Range =

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range =

Summary (*)

- Range
  - The maximum minus the minimum
  - Sensitive to outliers

- Standard deviation s
  - Measures deviations from the mean
  - Each value in the data set are used
  - Sensitive to outliers
  - Larger s implies greater variability

Activity: Descriptive Statistics

Consider the following three sets of observations:

- Set 1: 8,9,10,11,12
- Set 2: 8,9,10,11,100
- Set 3: 8,9,10,11,1000

(a) Find the mean and median for each data set.
(b) Find the range and standard deviation for each data set.
(c) What do these data sets illustrate about the resistance of the median and mean?

Use the online descriptive statistics tool to compute the answer
https://www.calculatorsoup.com/calculators/statistics/descriptivestatistics.php

Module1b - Descriptive

What is Probability?

- **Probability** is a measure of the likelihood that an event will occur.
- Probability is quantified as a number from 0 to 1.

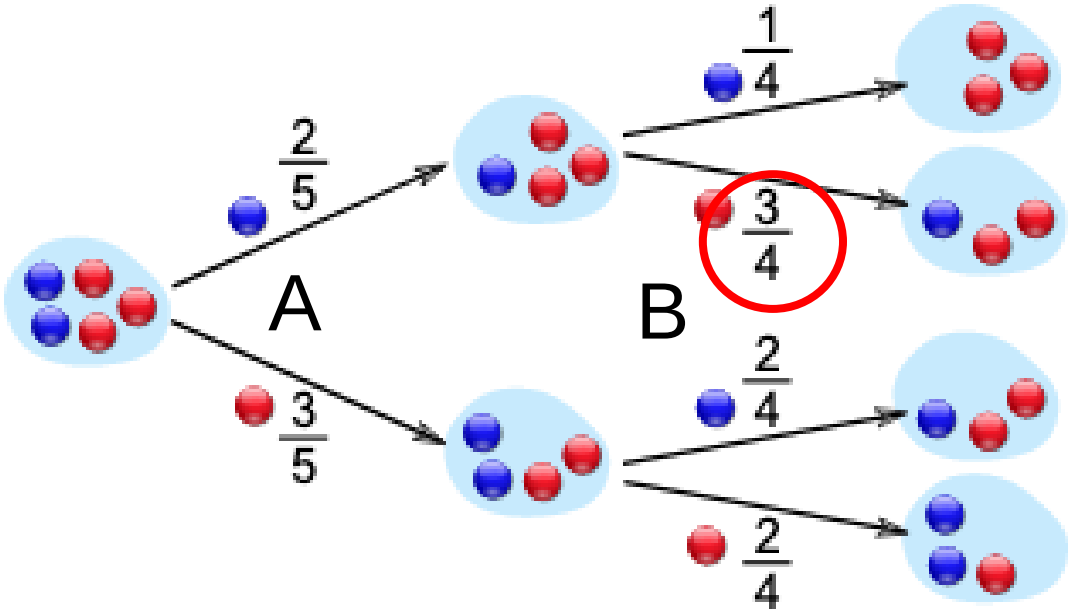
Probabilistic statements:

"The probability of getting a head from tossing a coin is 0.5"

"The probability of tomorrow rainy is high since today is a rainy day"

Conditional Probability

- Conditional probability is the probability of one event (B) given another event (A) is known/takes place
- A: get a blue marble first
- B: get a red marble then
- Based on the tree diagram below, P(B|A) = 3/4



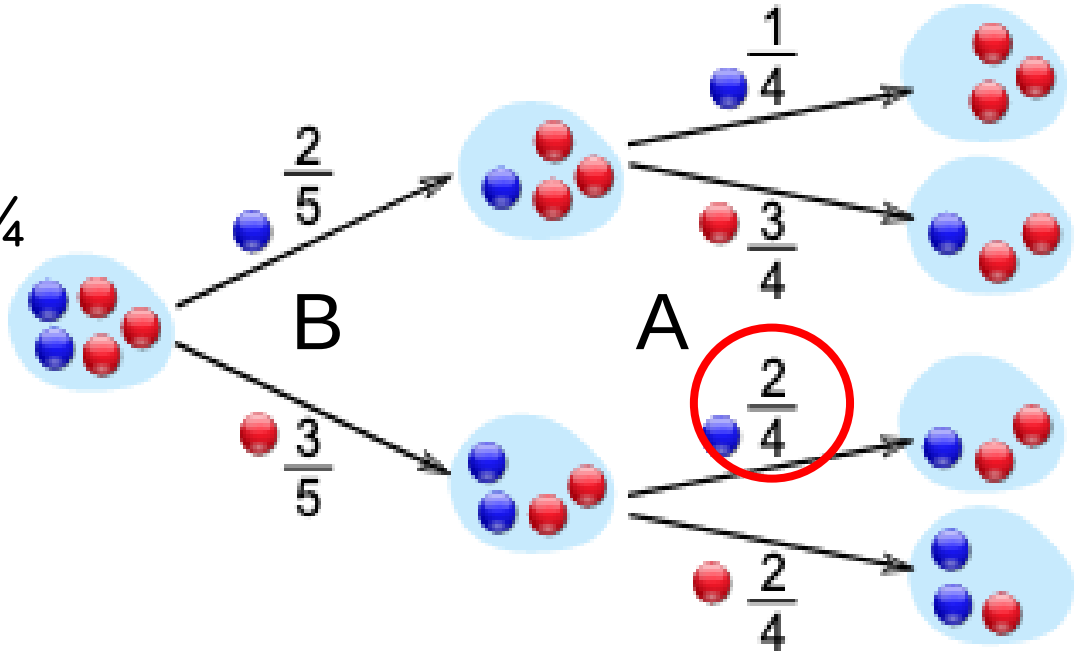
Bayes' Theorem

- Bayes' Theorem is a way of finding a conditional probability when we know other probabilities.
- The formula is P(A|B) = P(A) P(B|A)/P(B)

P(A|B) = P(B|A)*P(A)/P(B) = ¾ * ⅖ / ⅗ = ¾* ⅔ = 2/4

P(A|B) = P(A ∩ B) /P (B)

P(B|A) = ¾
P(A) = ⅖
P(B) = ⅗



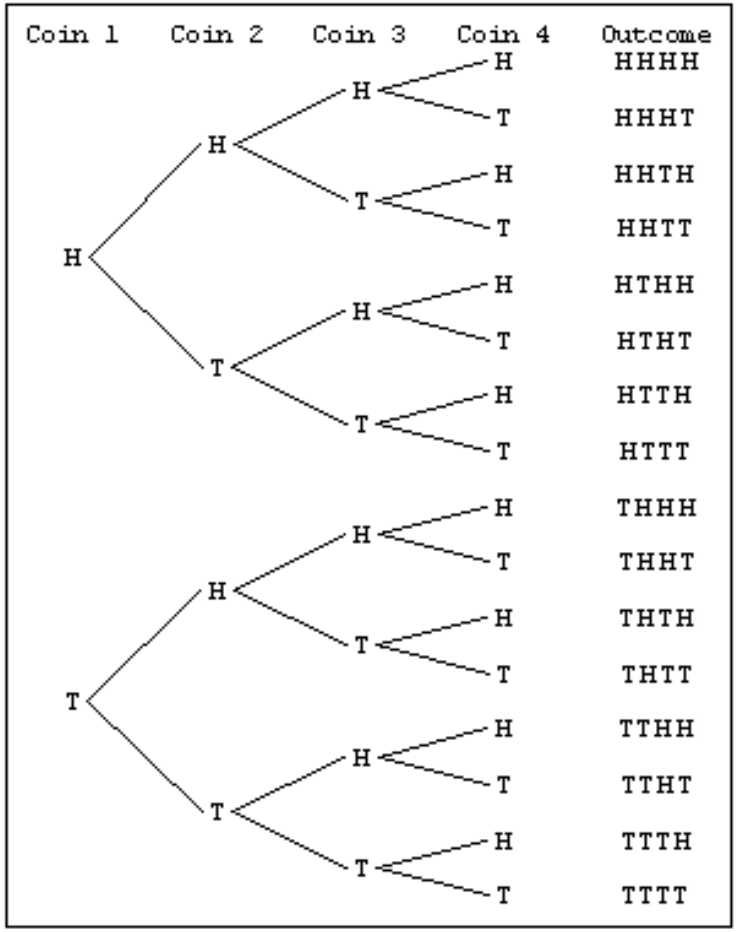
Activity: Conditional Probability

		Actual Status: Positive	Actual Status: Negative
Test Result	Positive	8 (TP)	2 (FP) Type 1 Error
	Negative	2 (FN) Type 2 Error	88 (TN)
		10	90

- If 100 people took the COVID-19 test, above is the test result.
- Compute the conditional probability that a person is positive is tested positive P(test positive|actual positive).

Binomial Distribution

- The binomial distribution with parameters n and ρ is the discrete probability distribution of the number of successes.
- A binomial distribution can be thought of as simply the probability of getting # of head outcome in an experiment of tossing multiple coins.

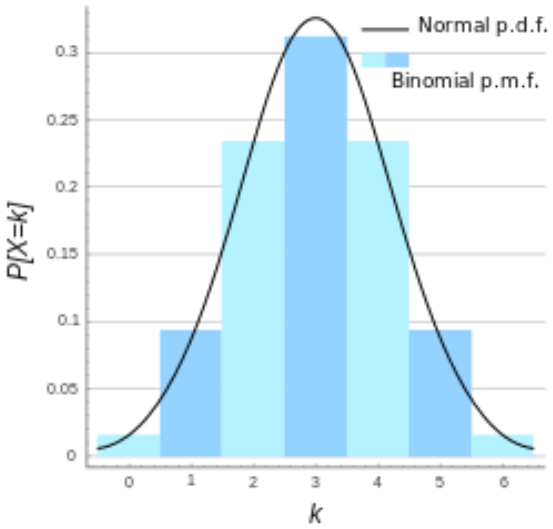


$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mu_X = np$$

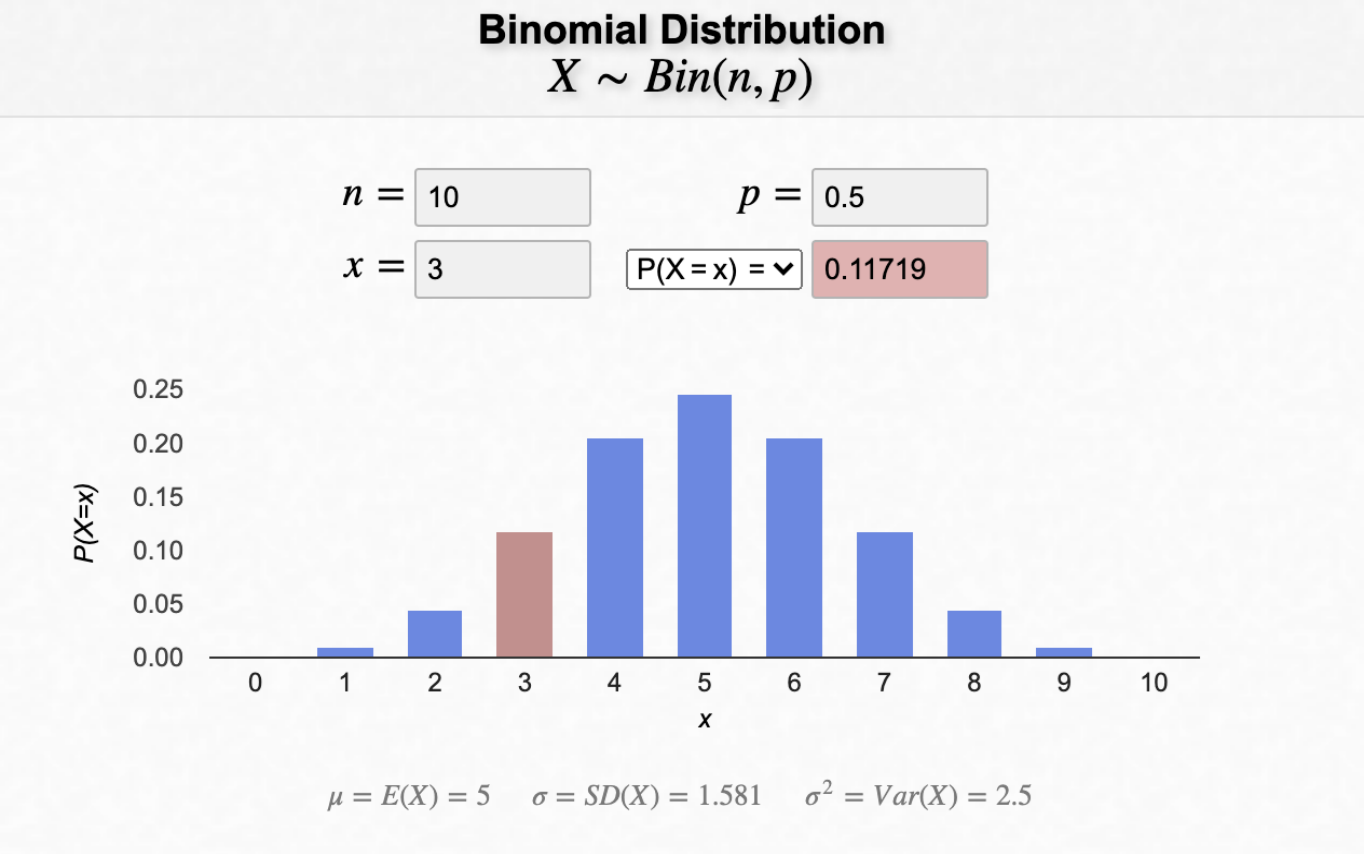
$$\sigma_X^2 = np(1-p)$$

n: No of coins/toss
k: No of head
ρ = Probability for getting a head in one toss



Activity: Binomial Distribution

- https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html
- Try out n= 10, p=0.5, x=3,4,5,6

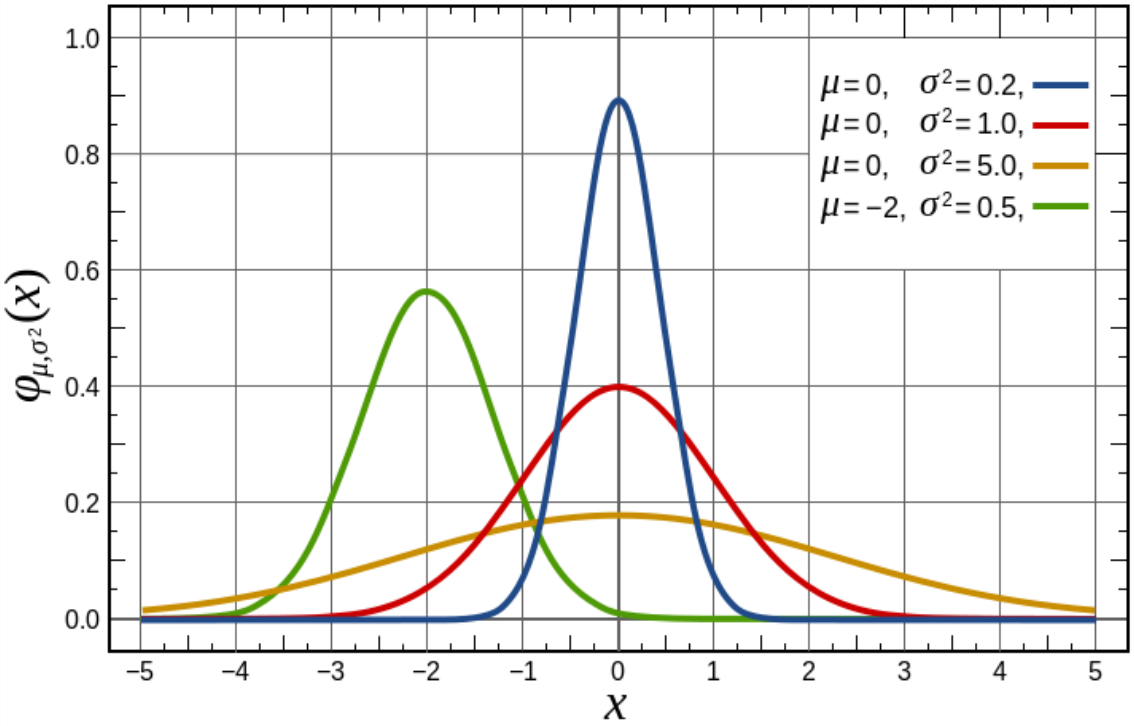


Module1c - Binomial

Normal Distribution

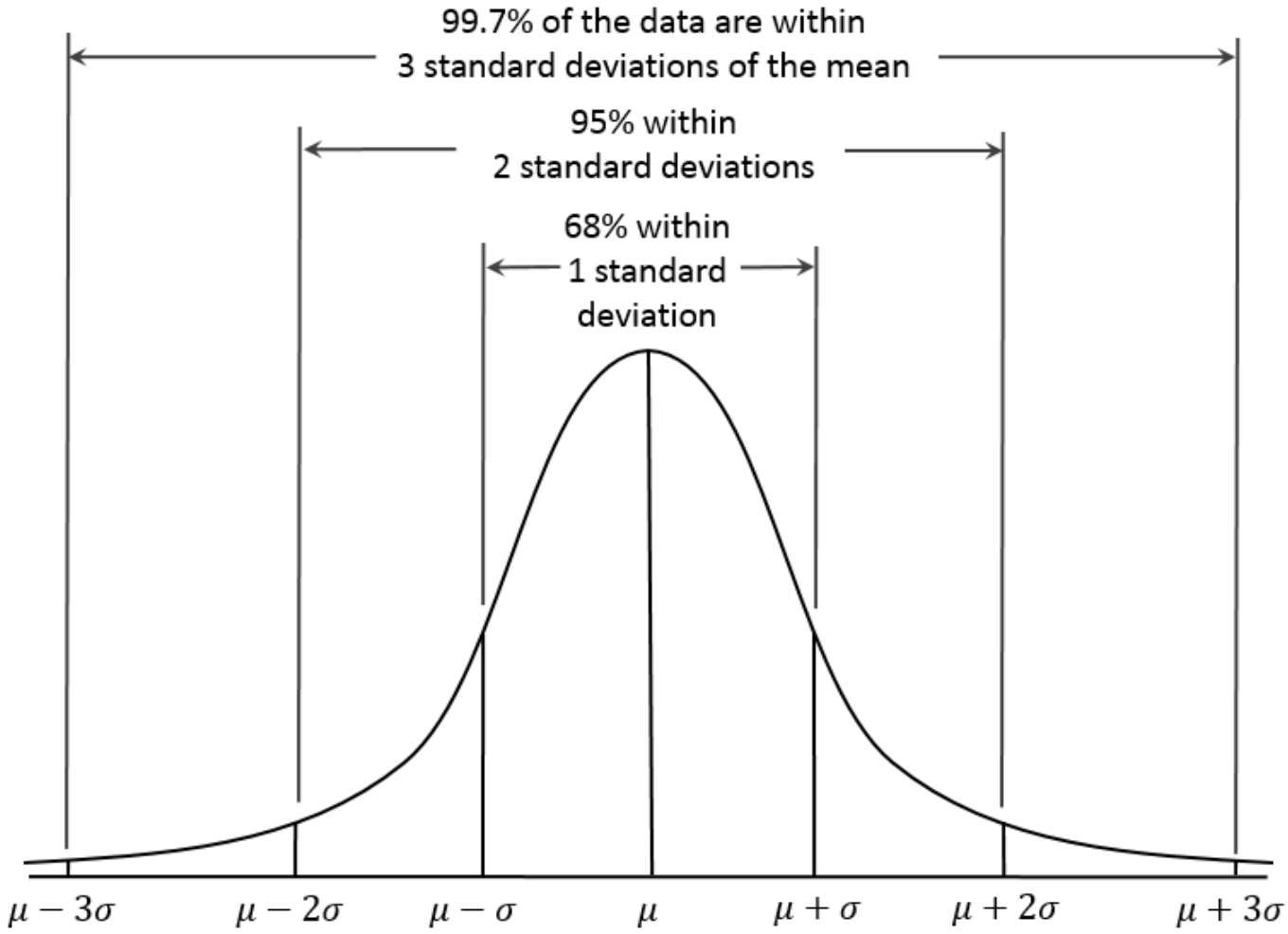
- Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is **symmetric** about the mean, showing that data near the mean are more frequent than data far from the mean.

- Normal distribution is approximation of Binomial distribution if n -> ∞.



$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normal Distribution (*)



Activity: Normal Distribution

- The normal distribution for women's' height in Singapore has μ= 160cm, σ= 20cm. Most major airlines have height requirements for flight attendants.

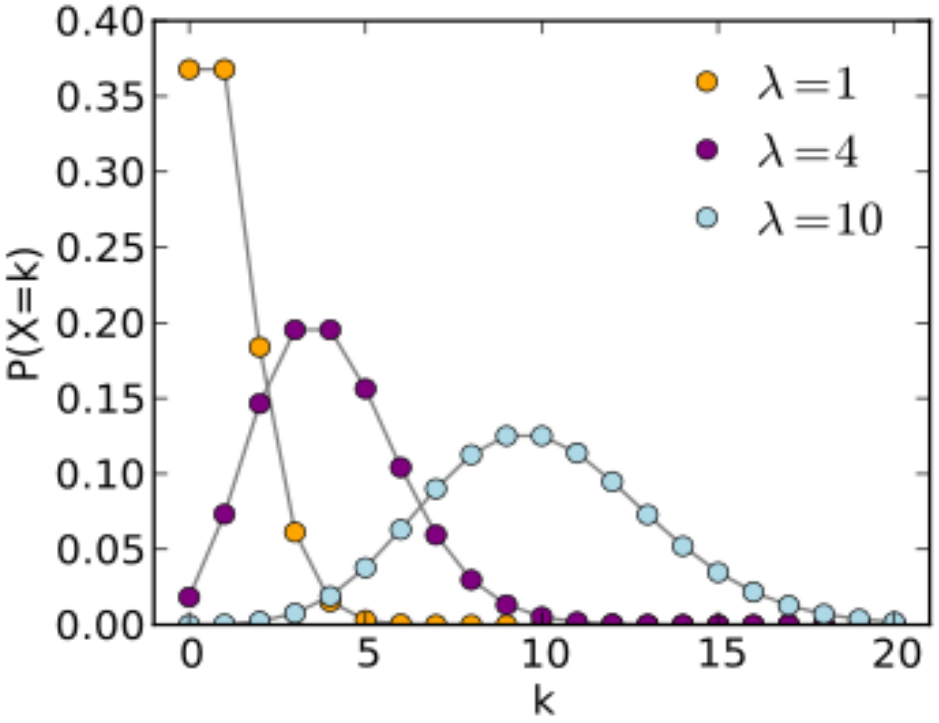
- The minimum height requirement is 170cm. What proportion of adult females in Singapore are not tall enough to be a flight attendant?

Use the following online Normal Distribution tool https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html to compute the answer

Module1d - Normal

Poisson Distribution

- The Poisson distribution lets you estimate the number of customers (events) who will come into a store during a given time period (fixed) such as an hour or perhaps the number of seconds between times that cars arrive at a toll booth.



$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

ere

What is Lambda (λ) and k? (*)

- λ, Lambda, or the rate parameter can be thought of as the expected number of events in the interval.

- k, how many times an event occurs in an interval.

Lambda (λ) is the total number of events (k) divided by the number of units (n, Time) in the data (λ = k/n)

Poisson Distribution Assumptions (*)

1. Events are **independent** of each other. The occurrence of one event does not affect the probability another event will occur.

2. The average rate (events per time period) (lambda) is constant.

3. Two events cannot occur at the same time.

Activity: Poisson Distribution

- A bank is interested in studying the number of people who use the ATM located outside its office late at night.

- On average, 1.6 customers use the ATM during any 10-minute interval between 9pm and midnight.

- What is lambda λ for this problem?

- What is the probability of exactly 3 customers using the ATM during any 10-minute interval?

- What is the probability of 3 or fewer people?

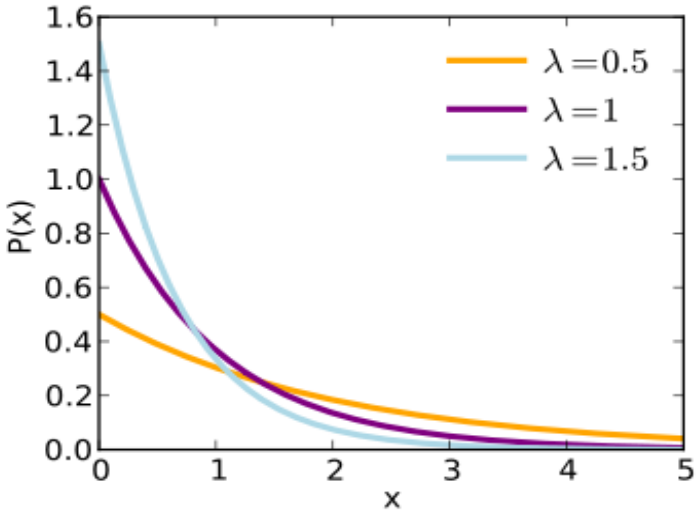
- Use the following Poisson online tool

https://homepage.divms.uiowa.edu/~mbognar/applets/pois.html

Module1e - Poisson

Exponential Distribution

- If you sell products via your company's website, knowing the average time between orders helps you plan the number of employees you'll need to have on duty at any time. It models the time we need to wait before a given event occurs.



$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Exponential Distribution Exs

- How much time will elapse before an earthquake occurs in each region?
- How long do we need to wait until a customer enters our shop?
- How long will it take before a call center receives the next phone call?
- How long will a piece of machinery work without breaking down?

This type of occurrence is described by the exponential probability distribution.

Calculating λ for Exponential Distribution (*)

The mean (μ) of an Exponential Distribution is $\frac{1}{\lambda}$, so λ is $\frac{1}{\mu}$

Activity: Exponential Distribution

- The number of days ahead travelers purchase their airline tickets can be modeled by an exponential distribution with the average amount of time equal to 15 days.

- Find the probability that a traveler will purchase a ticket fewer than ten days in advance.

- Use the exponential distribution online tool below

https://homepage.divms.uiowa.edu/~mbognar/applets/exp-like.html

Module1f - Exponential

Software for Statistics

- Minitab https://www.minitab.com/en-us/
- SPSS Statistics https://www.ibm.com/sg-en/products/spss-statistics
- Jamovi https://www.jamovi.org/ (Like a free version of SPSS with R code support)
- R https://cran.r-project.org/
- Excel Statistical Functions
- Python Data Analysis Packages like Pandas, Matplotlib and NumPy

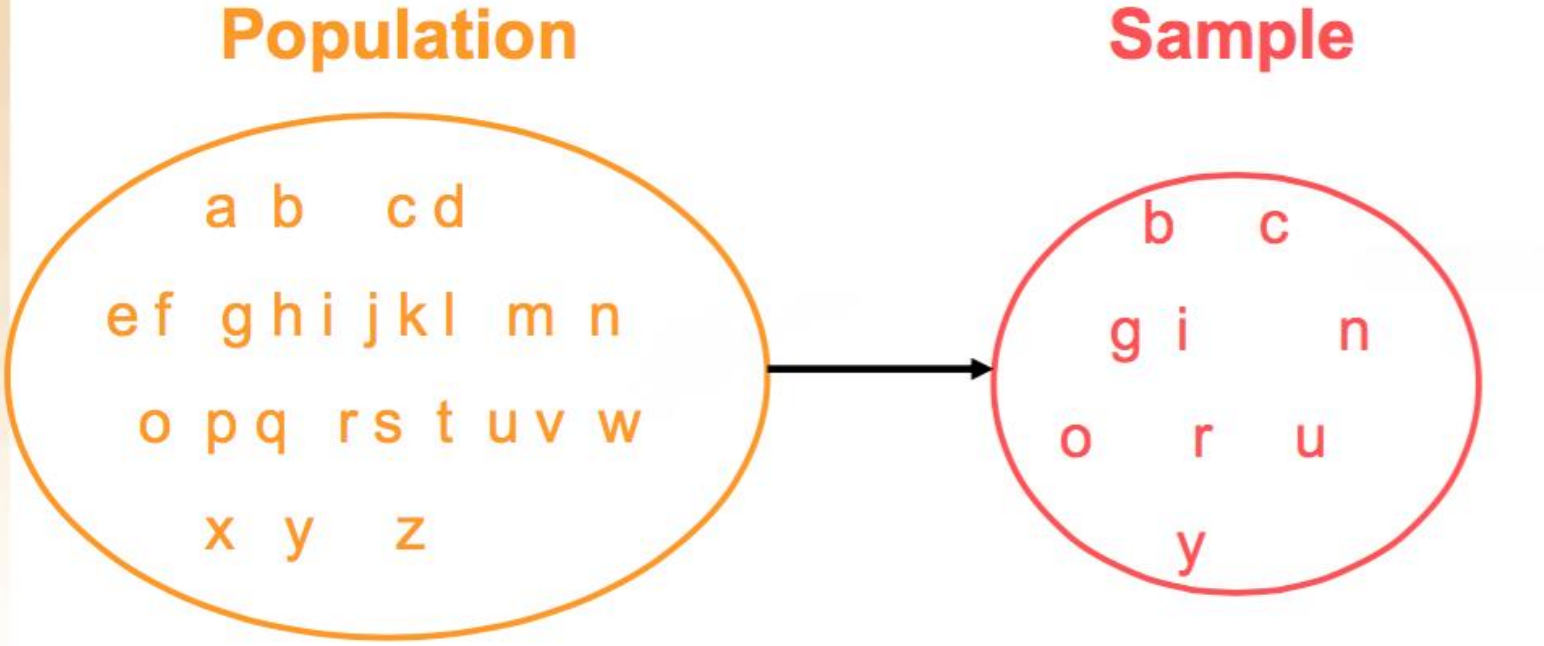
- Minitab, SPSS and Excel are commercial software.
- Jamovi and R are free open source software

Topic 2
Sampling

Sampling Theory

- Sampling theory is the field of statistics that is involved with the collection, analysis and interpretation of data gathered from random samples of a population under study.
- The application of sampling theory is concerned not only with the proper selection of observations from the population that will constitute the random sample
- It also involves the use of probability theory, along with prior knowledge about the population parameters, to analyze the data from the random sample and develop conclusions from the analysis.
- The normal distribution is most heavily utilized in developing the theoretical background for sampling theory.

Population vs Sample (*)



Measures used to describe a population are called parameters

Measures computed from sample data are called statistics

Term Definitions

- A **population** is the collection of all members of a group.
- A **sample** is a portion of the population selected for analysis.
- A **parameter** is a numerical measure that describes a characteristic of a population.
- A **statistic** is a numerical measure that describes a characteristic of a sample.

Sampling Distribution

- Parameters are usually unknown.
- Use statistics to estimate parameters.
- The ***sampling distribution*** of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.

Sample Mean & Standard Deviation

- The sample mean is a statistic that varies from sample to sample. - statistics, while population mean is a fixed value parameter.
- The estimate of the sample mean, and standard deviation is given below.
- Note that the n-1 instead of n in the sample standard deviation is to ensure an unbiased estimate of the popular standard deviation.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

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

Example: Pumpkin Weights

- The population is the weight of six pumpkins (in pounds) displayed in a carnival "guess the weight" game booth.

Pumpkin	A	B	C	D	E	F
Weight (in pounds)	19	14	15	9	10	17

Population Mean
=(19+14+15+9+10+17)/6=14 pounds

- You are asked to guess the average weight of the six pumpkins by taking a random sample without replacement from the population.

Example: Pumpkin Weights (sample size=2)

Sample	Weight	Sample mean
A, B	19, 14	16.5
A, C	19, 15	17.0
A, D	19, 9	14.0
A, E	19, 10	14.5
A, F	19, 17	18.0
B, C	14, 15	14.5
B, D	14, 9	11.5
B, E	14, 10	12
B, F	14, 17	15.5
C, D	15, 9	12
C, E	15, 10	12.5
C, F	15, 17	16
D, E	9, 10	9.5
D, F	9, 17	13
E, F	10, 17	13.5

- When using the sample mean to estimate the population mean, some possible error will be involved since sample mean is random.
- The chance that the sample mean is exactly the population mean is only 1/15 as the 15 different combinations of pumpkins.

Example: Pumpkin Weights (sample size=5)

Sample	Weight	Sample mean
A,B,C,D,E	19,14,15,9,10	13.4
A,B,C,D,F	19,14,15,9,17	14.8
A,B,C,E,F	19,14,15,10,17	15.0
A,B,D,E,F	19,14,9,10,17	13.8
A,C,D,E,F	19,15,9,10,17	14.0
B,C,D,E,F	14,15,9,10,17	13.0

- The chance that the sample mean is exactly the population mean is only 1/6.
- The error with a sample of size 5 is on the average smaller than with a sample of size 2.

Sampling Distribution of Sample Mean (*)

We can use StatKey to simulate the construction of a sampling distribution for a mean. You can access this simulation at:

http://www.lock5stat.com/StatKey/sampling_1_quant/sampling_1_quant.html

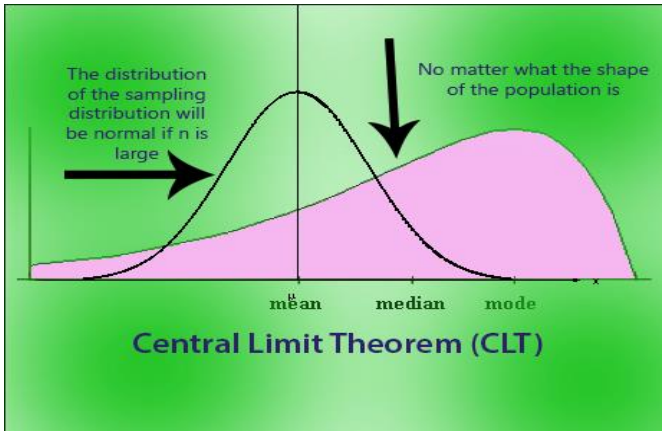
Central Limit Theorem

- For random sampling with a sample size n, the sampling distribution of the sample mean is approximately a normal distribution, no matter what the shape of the probability distribution from which the samples are taken.

- A rule of thumb for the sample size is more than 30. However, in (some) cases, a sample size of 5 or more can be sufficient.

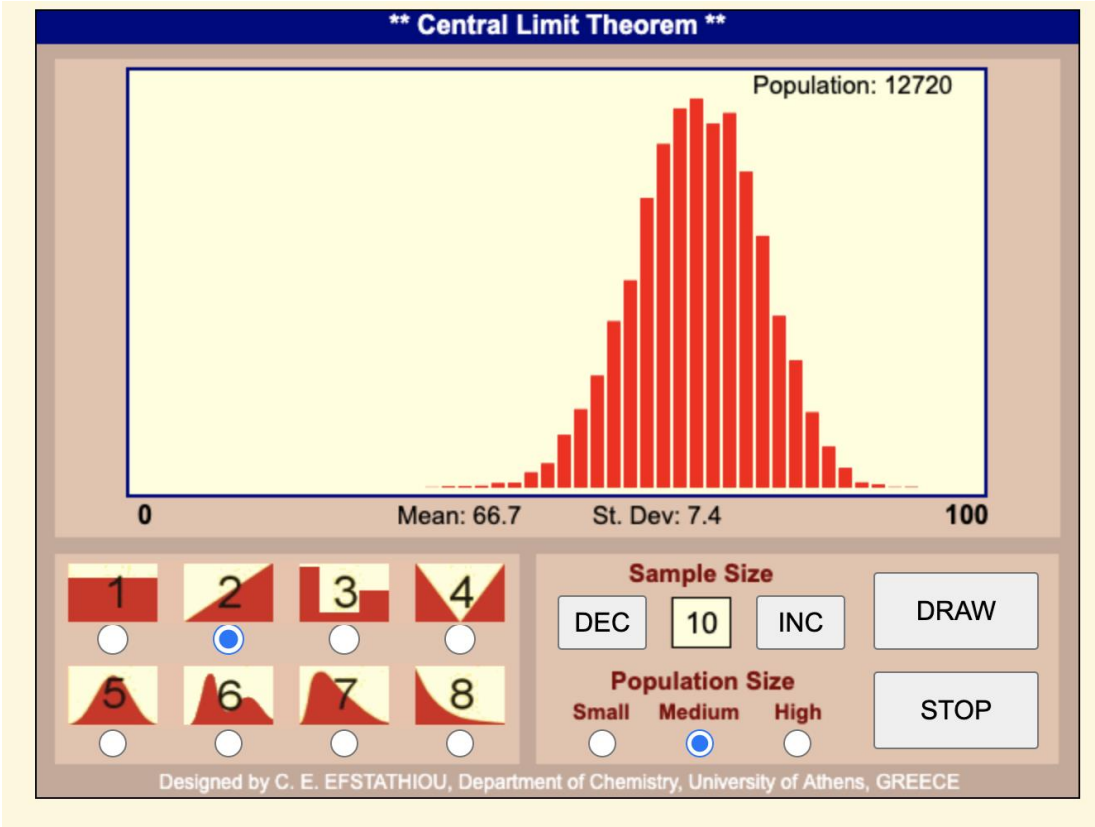
- The sample distribution standard deviation reduces as the sample size increases.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



Activity: Central Limit Theorem

- Try out eight different distributions, try a sample size of 5, 10,20, and see the sample mean distribution http://195.134.76.37/applets/AppletCentralLimit/Appl_CentralLimit2.html



Example (*)

Closing prices of stocks have a right skewed distribution with a mean of $25 and a standard deviation of $20.

Does the analysis meet the conditions of the CLT?

What is the probability that the mean of a random sample of 40 stocks will be less than $20?

Module2a – CLT Example

Applications of CLT

- Central Limit Theorem is used in several statistical areas such as :
  - Standard Error
  - Confidence Interval
  - Hypothesis Testing
  - ANOVA

Summary (*)

- Parameters are usually unknown.
- Use statistics to estimate parameters.
- Statistics are random.
- Sample distribution of sample mean: when X is normally distributed

$$\mathcal{N}(\mu, \sigma^2) \longrightarrow \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- CLT: when n>30, no matter what the shape of the population distribution

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Standard Error

- The standard error of a statistic is the standard deviation of the sampling statistic (mean, standard deviation, mode, median,..)

Sample 1
n, μ_1, σ_1,...

Sample 2
n, μ_2, σ_2,...

Sample 3
n, μ_3, σ_3,...

Sample 4
n, μ_4, σ_4,...

Sample 5
n, μ_5, σ_5,...

n: Sample size
N: No of samples

Standard Error (SE) of mean = $\frac{\sqrt{\Sigma \mu_i - \overline{\mu}}}{N}$

Standard Error (SE) of standard deviation = $\frac{\sqrt{\Sigma \sigma_i - \overline{\sigma}}}{N}$

Estimate Standard Error

- It is common to estimate the standard error from just one sample. One method is to use bootstrapping method (involves coding), another method is to compute the standard error based on Central Limit Theorem (CLT) as follows:

Step 1 The formula to find the sample mean

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

Step 2 Formula to estimate sample standard deviation

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

Step 3 Formula to estimate **standard error (SE) of mean**

$$\mathrm{SE}_{\mu_x} = \frac{s}{\sqrt{n}}$$

Activity: Standard Error

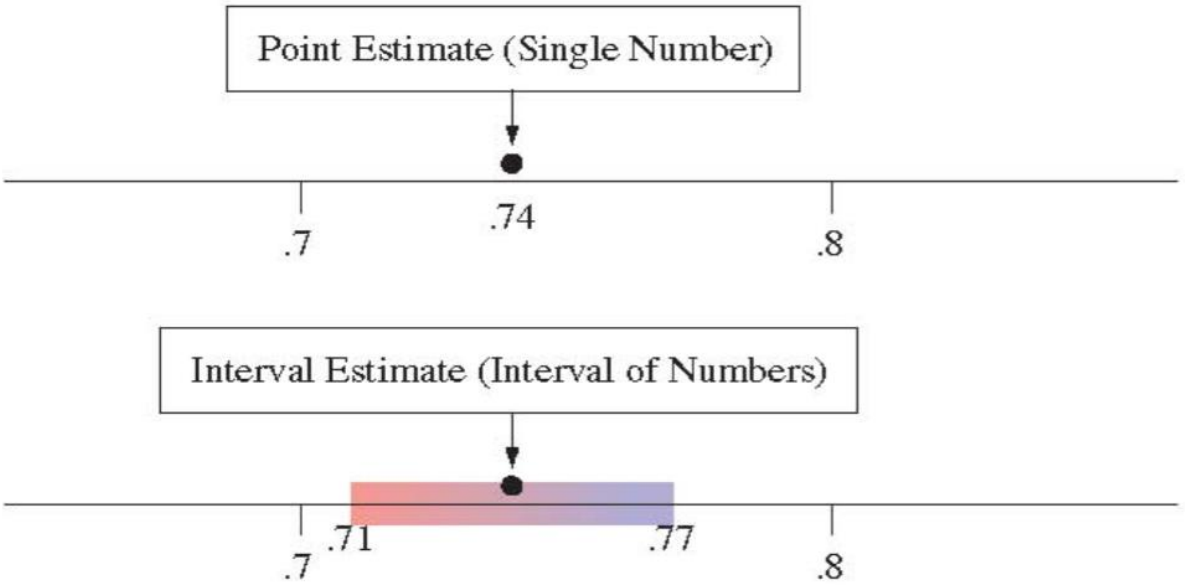
Consider the following three sets of observations:

- Set 1: 8,9,10,11,12
- Set 2: 8,9,10,11,100
- Set 3: 8,9,10,11,1000

Find the standard error using the following online tool:

https://ncalculators.com/statistics/standard-error-calculator.htm

Point Estimate vs Interval Estimate (*)



Point estimate: using a single number or point as the parameter estimate (0.74).

Interval estimate: the sample point estimate of 0.74 falls within a margin of error of 0.03

Point Estimate vs Interval Estimate (*)

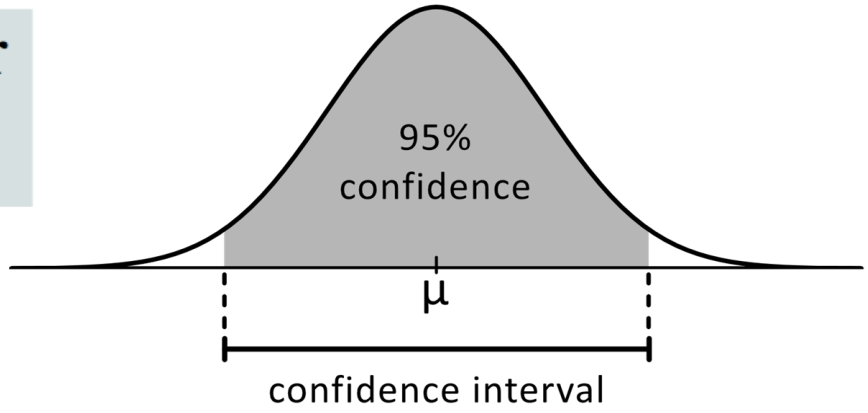
- A point estimate by itself is not sufficient because it doesn't tell us how close the estimate is likely to be to the parameter.
- An interval estimate is more useful. It incorporates a margin of error, so it helps us to gauge the accuracy of the point estimate.

Confidence Interval

- A confidence Interval is a range of values we are somewhat certain our true value lies in.
- This is a number chosen to be close to 1, most commonly 0.95 or 95% .
- When the sampling distribution is approximately normal, a 95% confidence interval has margin of error equal to 1.96 standard errors .

$$\text{point estimate} \pm \text{margin of error}$$

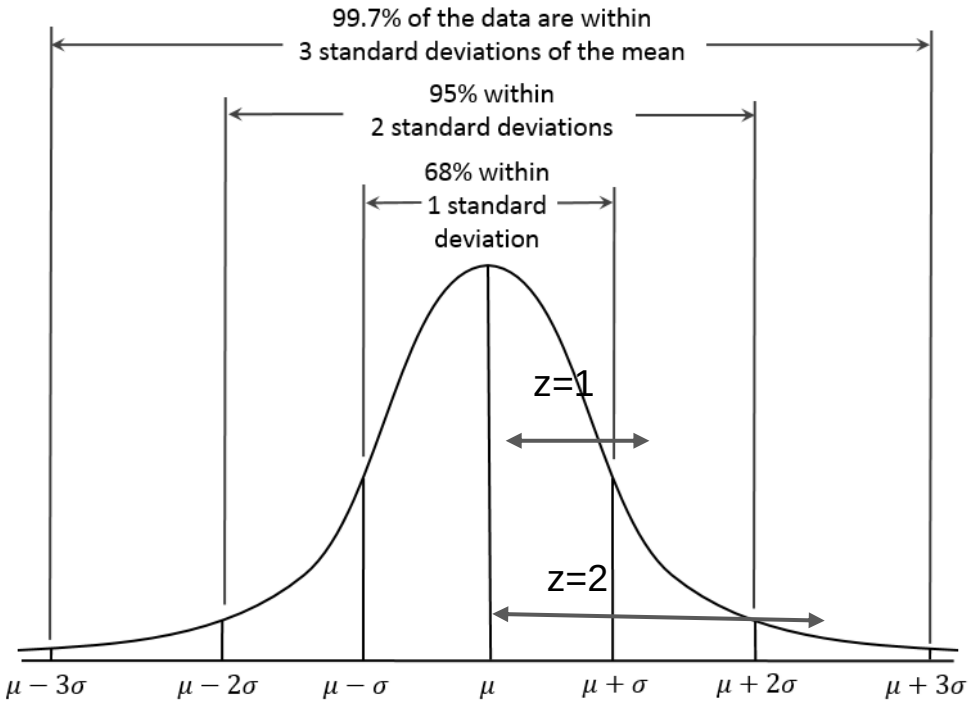
$$\text{point estimate} \pm 1.96(\text{se})$$



Z-Score

- The z-score for an observation is the number of standard deviations that it falls from the mean.
- The z-score is given by

$$z = \frac{(x - \mu)}{\sigma} \quad \text{(population)} \quad z = \frac{(\overline{x} - \mu)}{\sigma / \sqrt{N}} \quad \text{(sample)}$$



Activity: Z-Score

- Compute the z score (population) given
  - x = 154
  - μ = 100
  - $\boldsymbol{\sigma}$ = 30

You can use the online z-score calculator below
http://www.learningaboutelectronics.com/Articles/Z-score-calculator.php

Confidence Interval with Z-Score

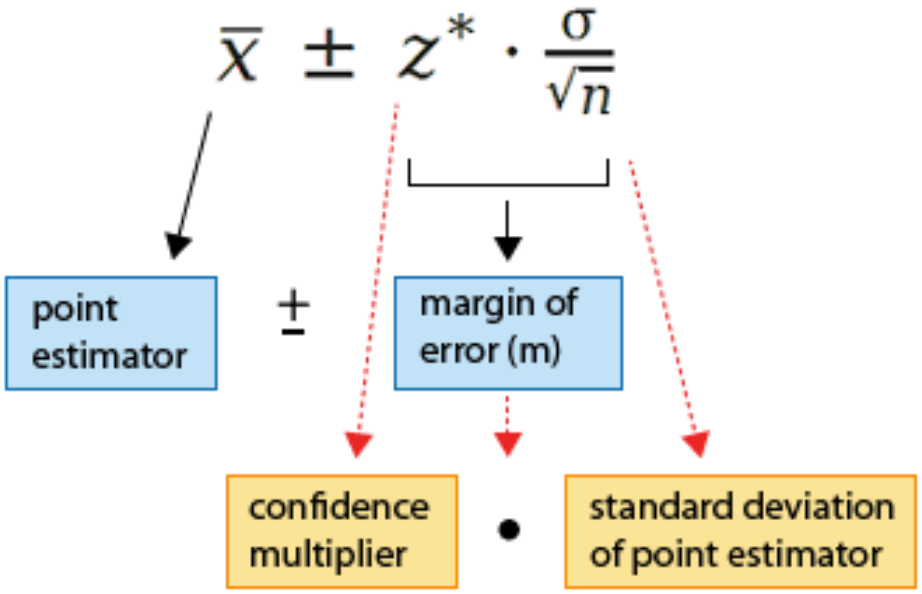
- The confidence interval for the population mean based on z-score, estimated from a sample for size n is:

$$\overline{x} \pm z \frac{\sigma}{\sqrt{N}}$$

Confidence level	Critical (z) value to be used in confidence interval calculation
50%	0.67449
75%	1.15035
90%	1.64485
95%	1.95996
97%	2.17009
99%	2.57583
99.9%	3.29053

CI for Population Mean (*)

For large n (n ≥30) or
For small n from a normal population



Activity: Confidence Interval (Z-Score)

- Suppose we know that the IQ scores of PSLE students are normally distributed with **population standard deviation of 15.**

- We have a simple random sample of 100 students, and the mean IQ score for this sample is 120.

- Find a 90% confidence interval for the mean IQ score for the entire population of all the PSLE students.

You can use the following confidence interval tools
https://www.mathsisfun.com/data/confidence-interval-calculator.html
https://www.socscistatistics.com/confidenceinterval/default3.aspx

Example (*)

Suppose you work for the Department of Natural Resources and you want to estimate, with 95% confidence, the mean (average) length of all walleye fingerlings in a fish hatchery pond.

Suppose you take a random sample of 100 fingerlings and determine that the average length is 7.5 inches; assume the population standard deviation is 2.3 inches.

Module2b – CI Example

Exercise 1 (*)

Suppose we know that the IQ scores of all incoming college freshman are **normally distributed with standard deviation of 15**.
We have a simple **random sample of 100** freshmen, and the mean IQ score for this sample is **120**.

Find a **90% confidence interval** for the mean IQ score for the entire population of incoming college freshmen.

Module2c - CI Exercise 1

Exercise 2 (*)

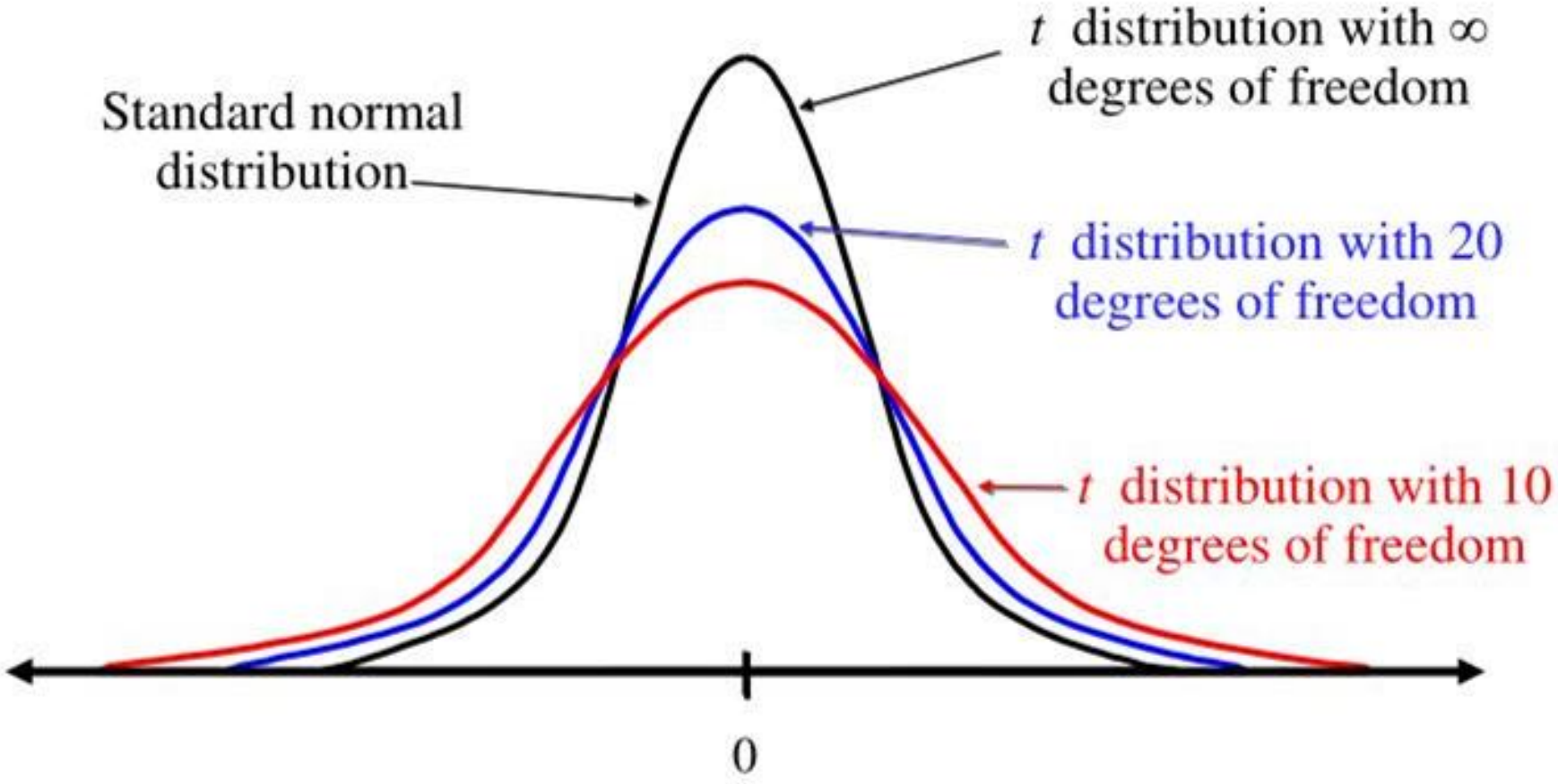
The 2012-2013 SASE scores of the **33 random** students from College of Science and Mathematics (CSM) of MSU-IIT were recorded: **84, 93, 101, 86, 82, 86, 88, 94, 89, 94, 93, 83, 95, 86, 94, 87, 91, 96, 89, 79, 99, 98, 81, 80, 88, 100, 90, 100, 81, 98, 87, 95, and 94.**
The population of these scores are believed to be **normally distributed with 6.8 standard deviation**.

Determine and interpret the **95% and 99% confidence interval** of the population mean.

Module2d - CI Exercise 2

Student's t Distribution

The t-distribution is used when n is **small** and σ is **unknown**.



Confidence Interval with T-Score

- You use the t score for small sample size (N<30), or unknown population standard deviation (very common)
- The t score is computed by

$$t = \frac{(\overline{x} - \mu)}{s/\sqrt{N}}$$

- Traditionally we look up a t values in a t-table. The number of items in your sample, minus one, is your degrees of freedom. For example, if you have 20 items in your sample, then df = 19.
- The confidence interval for the population mean based on t-score is:

$$\overline{x} \pm t\frac{s}{\sqrt{N}}$$

CI for Population Mean using t-score

Review CI for Population Mean using Z-score:

$$\overline{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

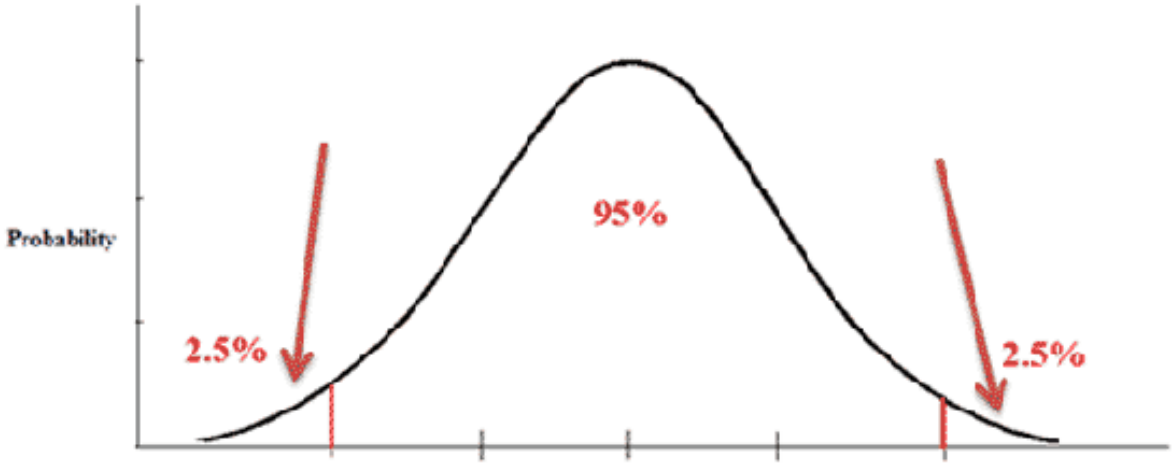
The CI for the population mean using t-score is:

$$\overline{\mathrm{X}} \pm t \frac{\mathrm{s}}{\sqrt{\mathrm{n}}}$$

s: sample standard deviation

95% CI for Population Mean

$$\bar{x} \pm t_{\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$$



$\alpha = 0.05$

$t_{0.05/2} = t_{0.025}$

Confidence Interval using T-score

Because the sample size is small, we need to use the t distribution.
For 95% confidence and df = n-1 = 9, t = 2,262.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792

Activity: T Value

- Verify the T-value for 95% confidence for a sample size of 10 is consistent with the T-value table using the following online t-value tool

http://www.learningaboutelectronics.com/Articles/T-value-calculator.php

Activity: Confidence Interval (T-Score)

- We have a small random sample of **10** students from the IQ scores of all PSLE students . The mean IQ score for this sample is 120 and sample **standard deviation** is 15.

- Find a 90% confidence interval for the mean IQ score for the entire population of PSLE students.

- You can use the following confidence interval tools based on t-score

https://www.socscistatistics.com/confidenceinterval/default2.aspx

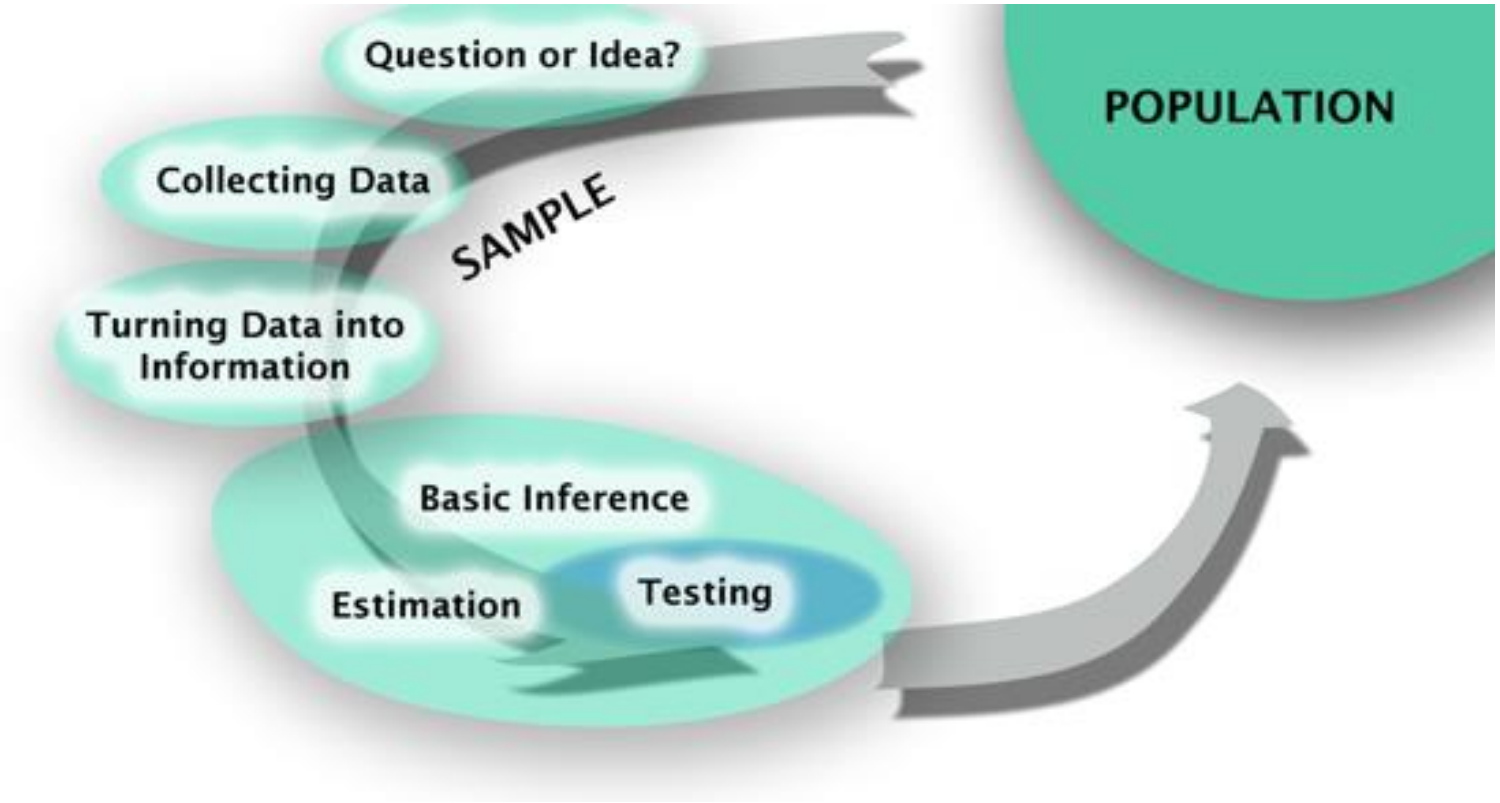
Exercise: Emergency Room Wait Time (*)

You are interested in the average emergency room (ER) wait time at your local hospital. You take a random sample of 50 patients who visit the ER over the past week. From this sample, the mean wait time was 30 minutes and the standard deviation was 20 minutes. Find a 95% confidence interval for the average ER wait time for the hospital.

Module2e - Exercise Emergency

Topic 3
Hypothesis Testing

Make Statistical Inferences about a Population (*)



What is Hypothesis Testing

- A **hypothesis** is a statement about a population, usually of the form that a certain parameter takes a particular numerical value or falls in a certain range of values .

- The main goal in many research studies is to check whether the data support certain hypotheses.

- A **hypothesis test** (significance test) is a method of using data to summarize the evidence about a hypothesis.

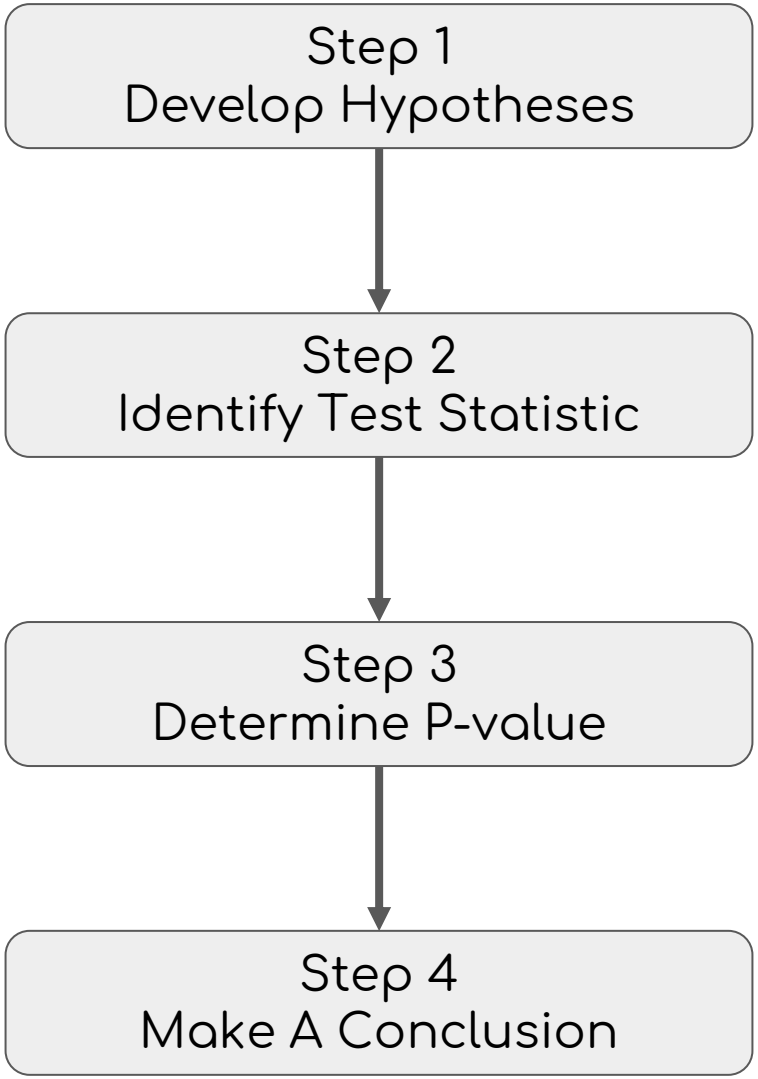
Example (*)

A principal at a certain school claims that the students in his school are above average intelligence.

A random sample of **30 students** IQ scores have a **mean score of 112**. Is there sufficient evidence to support the principal's claim?

The mean population IQ is 100 with a standard deviation of 15. IQ scores are normally distributed.

Steps of a Hypothesis Testing



Step 1: Develop Hypotheses

Each significance test has two hypotheses:

- The **null hypothesis (H_0)**states that
  - a parameter takes a particular value
  - a method has null effect (no effect).

- The **alternative hypothesis (H_a)** states that
  - a parameter falls in some alternative range of values.
  - a method has better or worst effect.
  - We usually *set the hypothesis* that **one wants to conclude** as the alternative hypothesis.

Examples of Hypotheses

Null Hypothesis	Alternative Hypothesis
Age has no effect on mathematical ability.	Mathematical ability depends on age
Taking aspirin daily does not affect heart attack risk.	Taking aspirin daily does affect heart attack risk.
Age has no effect on how cell phones are used for internet access.	Usage of cell phones for internet access depends on age
There is no difference in pain relief after chewing willow bark versus taking a placebo.	There is a difference in pain relief for chewing willow bark versus taking a placebo.

Exercise: $\mathbf{H_0}$ vs $\mathbf{H_a}$ (*)

Is the Statement a Null Hypothesis or an Alternative Hypothesis?

a) In Canada, the proportion of adults who favor legalized gambling is 0.50

a) The proportion of all Canadian college students who are regular smokers is less than 0.24, the value it was ten years ago.

H_0 vs H_a

Example: A consumer test agency wants to see the whether the mean lifetime of a brand of tires is less than 42,000 miles as the tire manufacturer advertises that the average lifetime is at least 42,000 miles.

H_0 vs H_a

Example: The length of a certain lumber from a national home building store is supposed to be 8.5 feet. A builder wants to check whether the shipment of lumber she receives has a mean length different from 8.5 feet.

Step 2: Identify Test Statistic

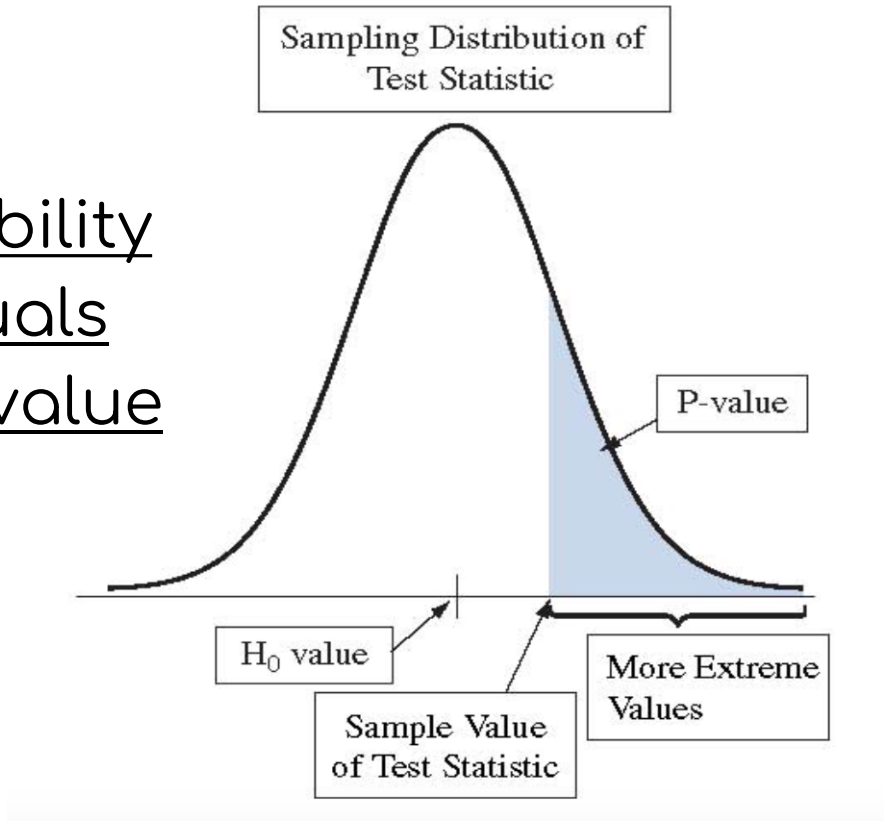
- A **test statistic** describes how far that estimate (the sample statistic) falls from the parameter value given in the null hypothesis.

- A test statistic is either z-score or t-score.

- In most cases, t-score is used as the sample size is small and the population variance is unknown (the more common reason).

Step 3: Determine P-value

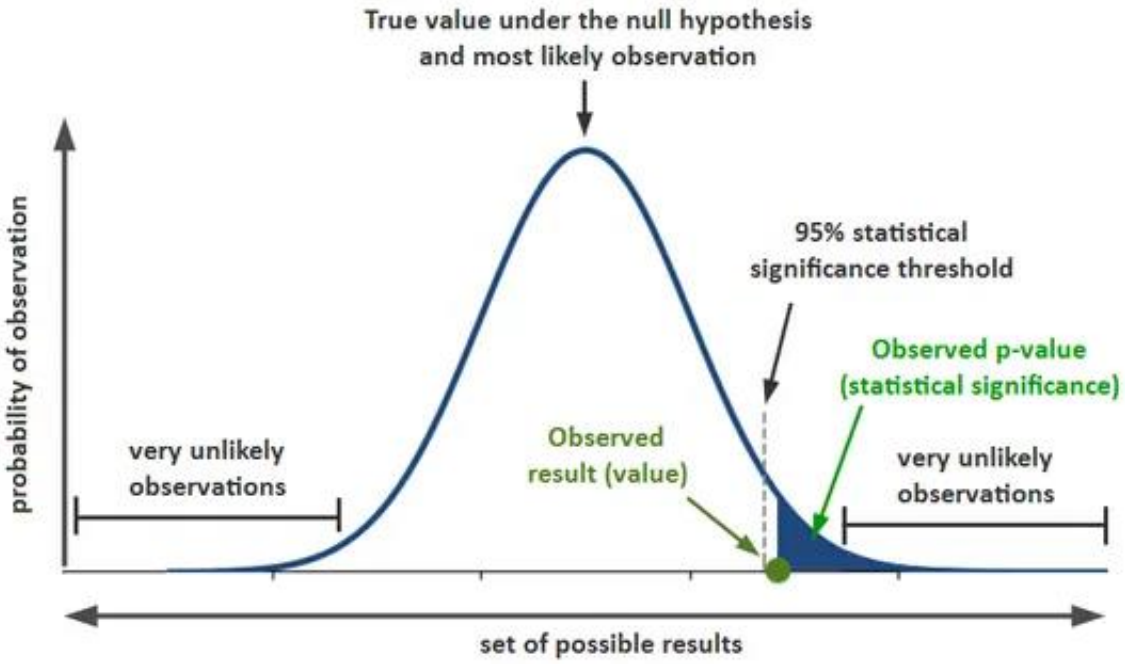
The p-value is the probability that the test statistic equals the observed value or a value even more extreme.



- The smaller the p-value, the stronger the evidence is against null hypothesis.

Significance Level

- The level of statistical significance is often expressed as a p-value between 0 and 1. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.
- A p-value less than 0.05 (typically ≤ 0.05) is statistically significant.



Significance Level

Significant Level	Specification
ρ>0.05	Not Significant
ρ<=0.05 (5%)	Significant
ρ<=0.01 (1%)	Very Significant
ρ<=0.001 (0.1%)	Highly Significant

- In practice, the most common significance level is 0.05

The alpha level, technically is arbitrary and without theoretical basis. The ρ-value can be 0.05000001 and we would not reject the null hypothesis.

Step 4: Make Conclusion

- Compare P-value with significance Level
- If the P-value < significance level, then reject the null hypothesis and accept the alternative hypothesis.

Note (*)

In hypothesis testing, we can never say we accept a hypothesis, we can only say we do not reject it and go with the alternative. A null hypothesis is not accepted just because it is not rejected.

Data not sufficient to show convincingly that a difference between means is not zero do not prove that the difference is zero. Such data may even suggest that the null hypothesis is false but not be strong enough to make a convincing case that the null hypothesis is false.

Type I and Type II Errors

Test Result	Actual Status: Positive (Alter)	Actual Status: Negative (Null)
Positive	(TP) True Positive Reject Null hypothesis	(FP) False Positive Reject Null hypothesis Type 1 Error
Negative	(FN) False Negative Accept Null Hypothesis Type 2 Error	(TN) True Negative Accept Null hypothesis

- Null Hypothesis => Negative
- Alternative Hypothesis => Positive

Example of Type I and Type II Errors

Type I error is more serious (*)

Set up the hypotheses so that Type I error is the more serious error.

Example : An inspector must choose between certifying a building as safe or saying that the building is not safe. There are two hypotheses:

1. Building is safe
2. Building is not safe

How will you set up the hypotheses?

Example: Building Inspections (*)

H_0: Building is not safe vs H_a: Building is safe

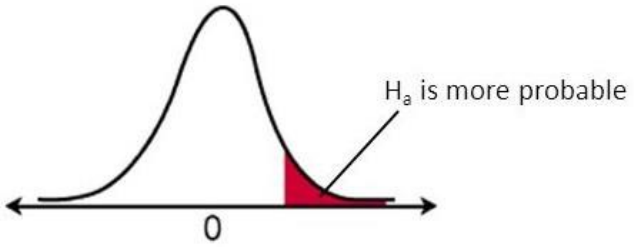
Decision	*Reality*	
	H_0 is true	H_0 is false
Reject H_0	Reject "building is not safe" when it is not safe (Type I error)	Correct
Fail to Reject H_0	Correct	Accept "building is not safe" when it is safe (Type II error)

Types of T-Tests

- **One sample t-test**: Compare a sample mean to a hypothetical mean.
- **One sample paired t-test**: Compare the difference from the same sample before and after treatment.
- **Two sample t-test**: Compare two sample means from two population with equal variances.
- **Two sample pooled variance t-test**: Compare two sample means from two population with unequal variances.
- It is recommended to use a two-sample t-test with unequal variance as we don't need to make assumption on the data.

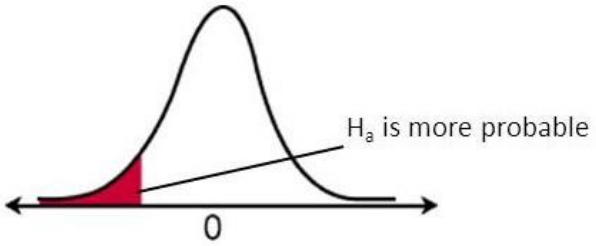
One Tail vs Two Tail T-Tests

- One tail t-test assumes the mean is higher or lower than a value.
- It is recommended to use a two-tail test as we don't need to make assumption on the data



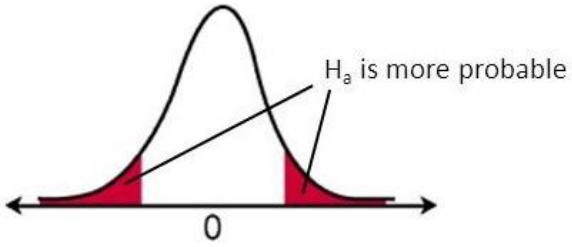
Right-tail test

$H_a: \mu > \text{value}$



Left-tail test

$H_a: \mu < \text{value}$



Two-tail test

$H_a: \mu \neq \text{value}$

Activity: One Sample Hypothesis Test

- We know that the national average (population) on the PSLE scoring system is 554 with a standard deviation of 99. Our sample of 90 students from ABC school had an average score of 568.

- Is the 14-point difference in averages enough to say that students in ABC school perform better than the national average at significance level 0.05?

- What is the ABC school average score is 590?

You can use the following too: http://www.learningaboutelectronics.com/Articles/Hypothesis-testing-calculator.php

Student IQ Example (*)

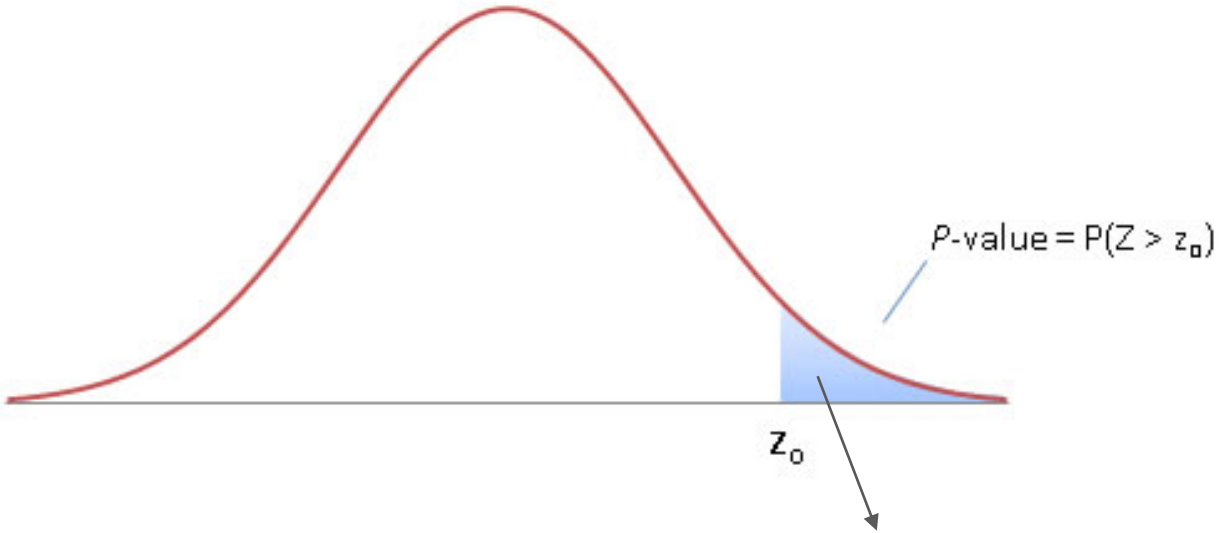
$H_0: \mu = 100$; The mean population IQ in his school is 100

$H_a: \mu > 100$; The students in his school have above average IQ scores.

Student IQ Example (*)

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{112 - 100}{15/\sqrt{30}} = 4.38$$

Student IQ Example (*)



4.38

P-value=0

Student IQ Example (*)

At significance level $\alpha=0.05$: p-value<0.05, thus we reject H_0. There is sufficient evidence to support the principal's claim.

Exercise 1 (*)

How do UMD students measure up on the older version of the verbal GRE? We know that the population average on the old version of the GRE (from ETS) was **554** with a standard deviation of **99**.

Our sample of **90** UMD students had an average of **568**. Is the 14-point difference in averages enough to say that UMD students perform better than the general population at significance level **0.05**?

Module3a - Exercise 1

Exercise 2 (*)

Suppose it is up to you to determine if a certain state (Michigan) receives a significantly different amount of public-school funding (per student) than the USA average.

You know that the USA mean public school yearly funding is $6800 per student per year, with a standard deviation of $400.

Suppose you collect a sample (n=100) from Michigan and determine that the sample mean for Michigan (per student per year) is $6873.

Exercise 2 (*)

Run a hypothesis test to determine if Michigan receives a significantly different amount of funding for public school education (per student per year).

Significance level=0.05

Module3b - Exercise 2

One Sample Paired T-test

- If you are studying the same group of students (one sample) before and after taking a special PSLE preparation session, you can use one sample paired t-test.

- You can use the following online for paired T-test http://www.learningaboutelectronics.com/Articles/Paired-t-test-calculator.php

Comparing Two Populations (Means) (*)

Example: We want to compare whether people give a higher taste rating to Coke or to Pepsi. To avoid possible psychological effect, the subjects should taste the drinks blind.

a) Randomly assign half of the subjects to taste Coke and the other half to taste Pepsi.

b) Allow all the subjects to rate both Coke and Pepsi.

Are the Data Independent Samples or Dependent Samples? (*)

- **Independent**: if the samples selected from one of the populations has no relationship with the samples selected from the other population.

- **Dependent** (Paired data): if each measurement in one sample is matched or paired with a particular measurement in the other sample.

Comparing Two Population Means: Independent Samples (*)

Example: There are two methods to purify water:

Method *A* : use a carbon filter
Method *B* : use a certain enzyme

μ1 = the mean bacteria count after using *A*;
μ2 = the mean bacteria count after using *B*.
To check whether μ1 =μ2 or μ1- μ2=0

Make an Inference about μ1- μ2 (*)

- Confidence Interval

  Point estimate ± margin of error

- Hypothesis Testing

  H_0 : μ1- μ2=0

CI for μ1- μ2 (*)

The point estimate for the difference is

$$\bar{X}_1 - \bar{X}_2$$

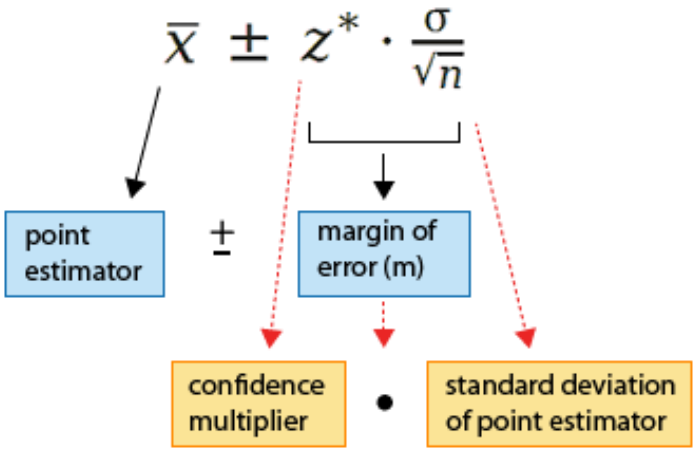
$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

In most cases, σ_1 and σ_2 are unknown so we must estimate σ_1 by s_1 and σ_2 by s_2

CI for μ1- μ2 (*)

CI of μ1- μ2:

$$(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

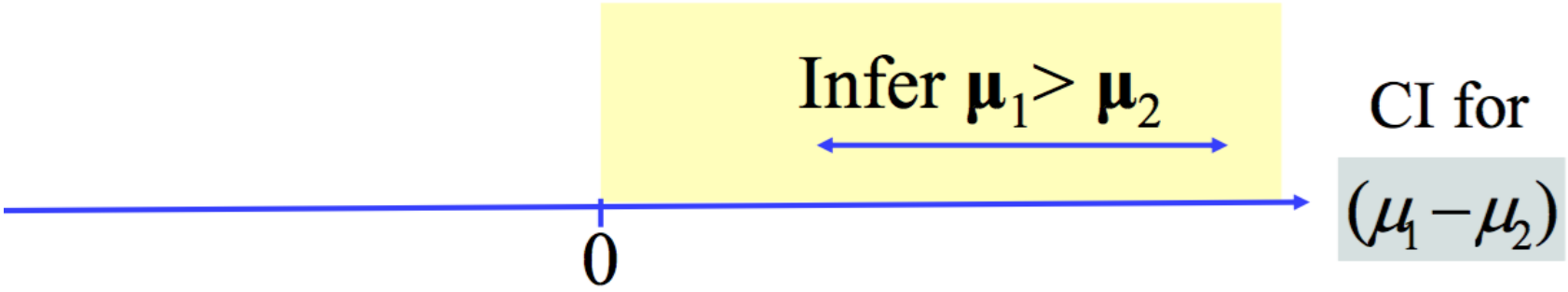


$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Interpreting a CI for μ1- μ2 (*)

A confidence interval for ($\mu_1 - \mu_2$) that contains only positive numbers suggests that ($\mu_1 - \mu_2$) is positive

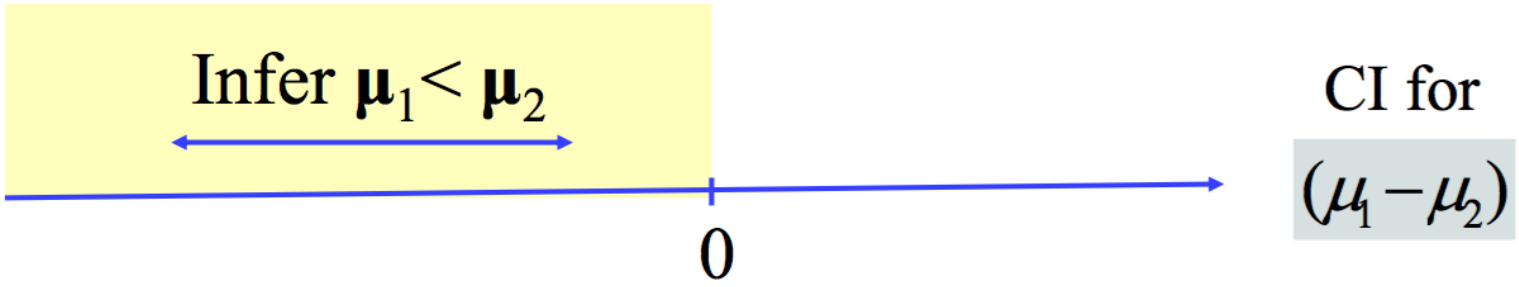
We then infer that μ_1 is larger than μ_2



Interpreting a CI for μ1- μ2 (*)

- **A confidence interval for ($\mu_1 - \mu_2$) that contains only negative numbers suggests that ($\mu_1 - \mu_2$) is negative**

  We then infer that μ_1 is smaller than μ_2



Example (*)

Do women tend to spend more time on housework than men? If so, how much more?

	Housework Hours		
Gender	**Sample Size**	**Mean**	**Standard Deviation**
Women	476	33.0	21.9
Men	496	19.9	14.6

Example (*)

a) Based on this study, calculate how many more hours, on the average, women spend on housework than men.

b) Find the standard error for comparing the means. What factors causes the standard error to be small compared to the sample standard deviations for the two groups?

c) Calculate the 95% CI comparing the population means for women and men. Interpret the result including the relevance of 0 being within the interval or not.

Module3c – Example Housework

Exercise (*)

A random sample of 100 students from MBA class made an average score of 60 with a standard deviation score of 15 in statistics.
A random sample of 64 students from BS class made an average score of 66 with a standard deviation of 16 in the same course. Construct a 95% confidence interval for the difference between the mean score of the two classes. df=128

2-sample t-test

The null hypothesis is the hypothesis of no difference or no effect:

$$H_0: (\mu_1 - \mu_2) = 0$$

The alternative hypothesis:

$$H_a: (\mu_1 - \mu_2) \neq 0 \text{ (two-sided test)}$$
$$H_a: (\mu_1 - \mu_2) < 0 \text{ (one-sided test)}$$
$$H_a: (\mu_1 - \mu_2) > 0 \text{ (one-sided test)}$$

Example (*)

Does Cell Phone Use While Driving Impair Reaction Times?

- 64 college students. 32 were randomly assigned to the cell phone group and 32 to the control group students used a machine that simulated driving situations
- At irregular periods, a target flashed red or green

Example (*)

- Participants were instructed to press a “brake button” as soon as possible when they detected a red light

- For each subject, the experiment analyzed their mean response time over all the trials

Example (*)

- Averaged over all trials and subjects, the *mean* response time for the cell-phone group was 585.2 milliseconds.

- The *mean* response time for the control group was 533.7 milliseconds.

Sample	N	Mean	StDev
Cell Phone	32	585.2	89.6
Control	32	533.7	65.3

Exercise (*)

Independent random samples of 17 sophomores and 13 juniors attending a large university yield the following data on grade point averages

Sophomores			Juniors		
3.04	2.92	2.86	2.56	3.47	2.65
1.71	3.60	3.49	2.77	3.26	3.00
3.30	2.28	3.11	2.70	3.20	3.39
2.88	2.82	2.13	3.00	3.19	2.58
2.11	3.03	3.27	2.98		
2.60	3.13				

Exercise (*)

At the 5% significance level, do the data provide sufficient evidence to conclude that the mean GPAs of sophomores and juniors at the university differ? (df=26)

(Start using ToolPak – t-Test assuming unequal variances)

Module3f - GPA

Two Samples Pooled Variance T-Test

- For two samples from two populations with different variances, the null hypothesis is given by:

$$H_o = \mu_1 - \mu_2$$

- The t-test statistics is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with degrees of freedom equal to n_1+n_2-2

Exercise (*)

In a packing plant, a machine packs cartons with jars. It is supposed that a new machine will pack faster on the average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded.

Exercise (*)

New machine					Old machine				
42.1	41.3	42.4	43.2	41.8	42.7	43.8	42.5	43.1	44.0
41.0	41.8	42.8	42.3	42.7	43.6	43.3	43.5	41.7	44.1
$\bar{y}_1 = 42.14, s_1 = 0.683$					$\bar{y}_2 = 43.23, s_2 = 0.750$				

Do the data provide sufficient evidence to conclude that, on the average, the new machine packs faster?

Module3g - Machine

Activity: Two Samples Hypothesis Test

- Do women tend to spend more time on housework than men? If so, how much more?

Housework Hours			
Gender	Sample Size	Mean	Standard Deviation
Women	476	33.0	21.9
Men	496	19.9	14.6

You can analyze two samples hypothesis testing with 5% significant level using the tool below
http://www.statskingdom.com/140MeanT2eq.html

Activity: Two Samples Hypothesis Test

- Independent random samples of 17 students from JC 1 and 13 students from JC 2 yield the following grade points.
- Is there any difference in grade points between JC 1 and JC 2 students?

JC 1			JC 2		
3.04	2.92	2.86	2.56	3.47	2.65
1.71	3.60	3.49	2.77	3.26	3.00
3.30	2.28	3.49	2.70	3.20	3.39
2.88	2.82	2.13	3.00	3.19	2.58
2.11	3.03	3.27	2.98		
2.60	3.13				

You can analyse using two samples hypothesis testing with 5% significance level using the tool below
https://www.socscistatistics.com/tests/studentttest/default2.aspx

Topic 4
Chi Square Test

M&M candies Example (*)

There are six different colors: red, orange, yellow, green, blue and brown. Suppose that we are curious about the distribution of these colors and ask, do all six colors occur in equal proportion?

M&M candies Example (*)

Suppose that we have a simple random sample of 600 M&M candies with the following distribution:

Blue	212
Orange	147
Green	103
Red	50
Yellow	46
Brown	42

What is Chi-Square Testing

- The Chi Square statistic is commonly used for testing relationships between categorical variables.

- There are two main kinds of chi-square tests: **test of independence** and **goodness-of-fit test**.

- Test of Independence asks a question of relationship, such as, "Is there a relationship between gender and A level scores?";

- Goodness-of-fit test asks something like "If a coin is tossed 100 times, will it come up heads 50 times and tails 50 times?"

Why not t test? (*)

T-test: Quantitative data

Chi-squared test: Categorical Data

Goodness Fit Test (*)

- The goodness of fit of a statistical model describes how well it fits a set of observations.

- Measures of goodness of fit typically summarize the discrepancy between observed values and the expected values under the model in question.

Goodness Fit Test (*)

Step 1: Set Up Hypothesis

Null hypothesis

H_0 = The stated distribution is accurate.

Alternative hypothesis

H_a = The stated hypothesis is not accurate.

M&M candies Example (*)

Null hypothesis: the population proportion of all colors are equal, that is ⅙.

Alternative hypothesis: at least one of the population proportions is not equal to 1/6.

Goodness Fit Test (*)

Step 2: Calculate Chi-Squared Statistics

$$X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

M&M candies Example (*)

If the null hypothesis were true, then the expected counts for each of these colors would be (1/6) x 600 = 100.

Chi-Square Statistic

- The steps for chi-square hypothesis testing is like hypothesis testing using t-statistics.
- The chi-square statistics is given by:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:
c=Degrees of freedom
O=Observed value(s)
E=Expected value(s)

- A very small chi-square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.
- A very large chi-square test statistic means that the data does not fit very well. In other words, there isn't a relationship.

M&M candies Example (*)

Color	Observed	Expected	$(O- E)^2/E$
Blue	212	100	$(212 – 100)^2/100 = 125.44$
Orange	147	100	$(147 – 100)^2/100 = 22.09$
Green	103	100	$(103 – 100)^2/100 = 0.09$
Red	50	100	$(50 – 100)^2/100 = 25$
Yellow	46	100	$(46 – 100)^2/100 = 29.16$
Brown	42	100	$(42 – 100)^2/100 = 33.64$

125.44 + 22.09 + 0.09 + 25 +29.16 + 33.64 =235.42

Degrees of Freedom (*)

df=k-1

where k is the number of categories.

M&M candies Example: k=6, so df=6-1=5

Chi-Square Statistic Example

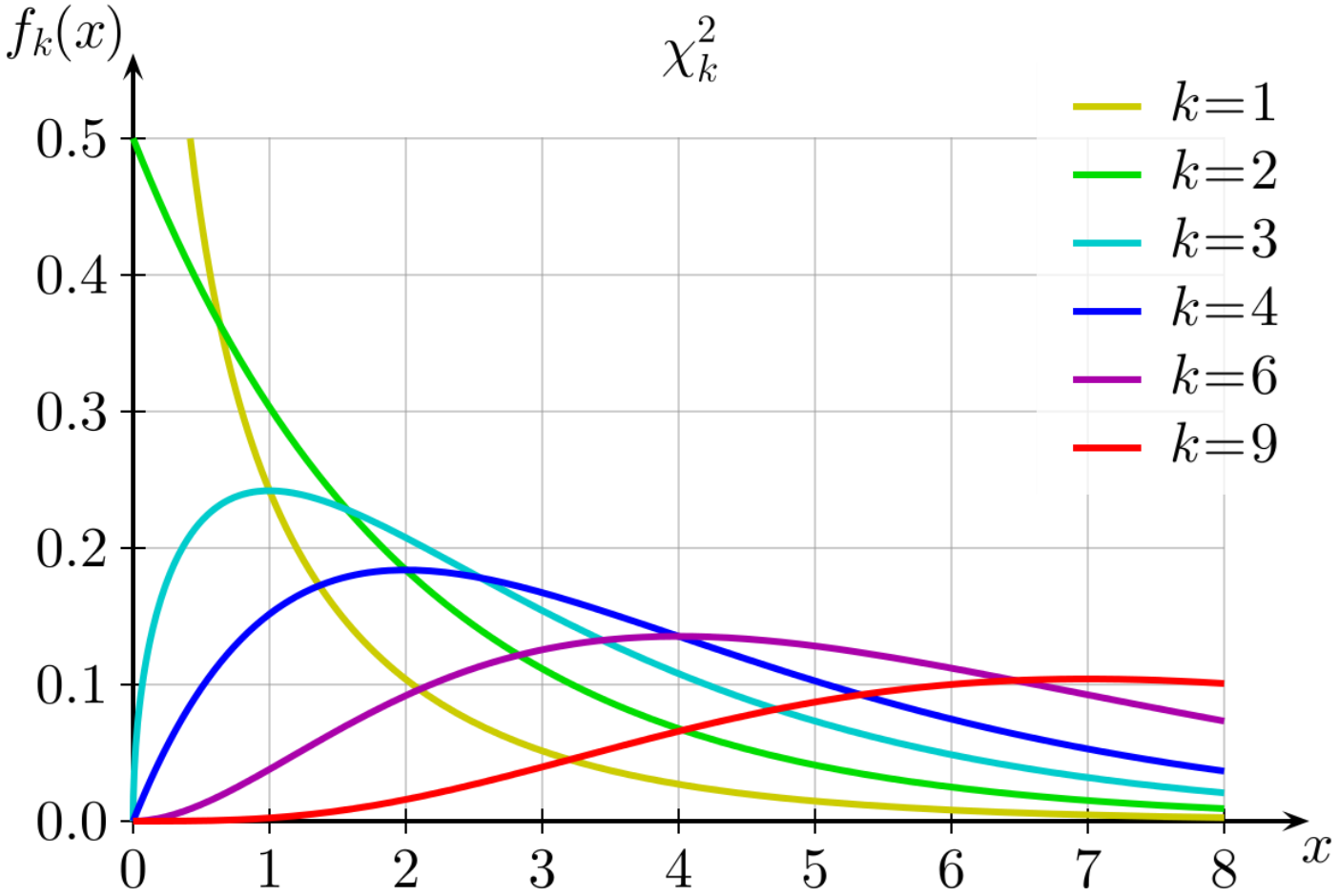
- 10 coins - if these 10 coins are, they fair.
- If fair coin, 50% head, 50% tail for each coin

	Observation (H)	Expectation (H)
H	7	5
T	3	5

$$\chi^2 = \frac{(7-5)^2 + (3-5)^2}{5} = \frac{4+4}{5} = 1.6$$

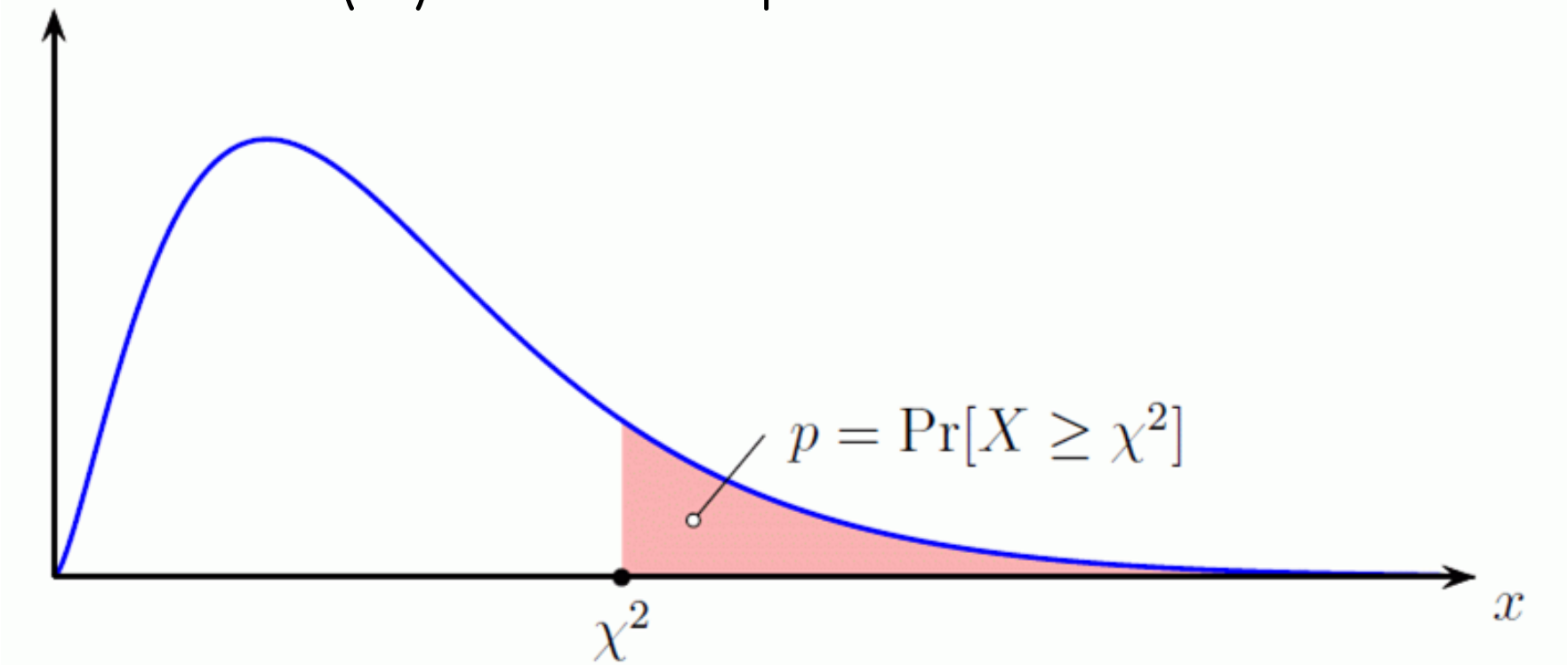
Chi-Squared Distribution

For each degree of freedom, we have a different chi-squared curve



P-Value of Chi Square Statistic

The P-value of Chi Square depends on the degree of freedom (df) and Chi Square score



You can try out the following tool to compute the p-value for chis square
http://courses.atlas.illinois.edu/spring2016/STAT/STAT200/pchisq.html

Activity: Chi Square Test

- For a fair six-sided die, the probability of any given outcome on a single roll would be 1/6. The data in Table were obtained by rolling a six-sided die 36 times. Are these data consistent with the hypothesis that the die is a fair die?

Dice	Frequency
1	8
2	5
3	9
4	2
5	7
6	5

You can use the following tool for Chi Square Test
https://www.graphpad.com/quickcalcs/chisquared1/

Dice Example (*)

P-value= **1-CHISQ.DIST**(1.6,1,TRUE)=0.206,
so we cannot reject the null hypothesis. We conclude that the dice is fair.

M&M candies Example (*)

P-value= **1-CHISQ.DIST**(235.42,5,TRUE) = 0,
so we reject null hypothesis. We conclude that M&Ms are not evenly distributed among the six different colors.

Exercise 1 (*)

For a fair six-sided die, the probability of any given outcome on a single roll would be 1/6. The data in Table were obtained by rolling a six-sided die 36 times. Are these data consistent with the hypothesis that the die is a fair die?

Outcome	Frequency
1	8
2	5
3	9
4	2
5	7
6	5

Exercise 2 (*)

A safari park in Africa is divided into 8 zones, each containing a known population of elephants. A sample is taken of the number of elephants found in each zone to determine whether the distribution of elephants is significantly different from what would be expected based on the known population in each zone.

Exercise 2 (*)

Hint: the expected number of elephants in Zone 1 should be 24/205x55 based on the known population

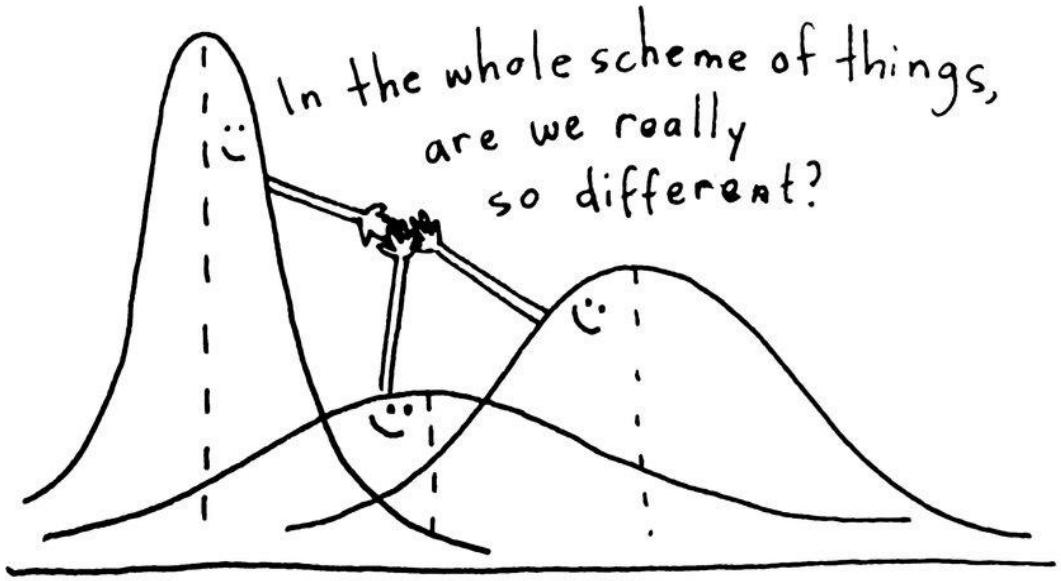
Zone	Sample	Population
1	7	24
2	11	27
3	5	20
4	6	22
5	9	35
6	4	17
7	5	32
8	8	28
Total	55	205

Topic 5
ANOVA: Analysis of Variance

Analysis of Variance (ANOVA)

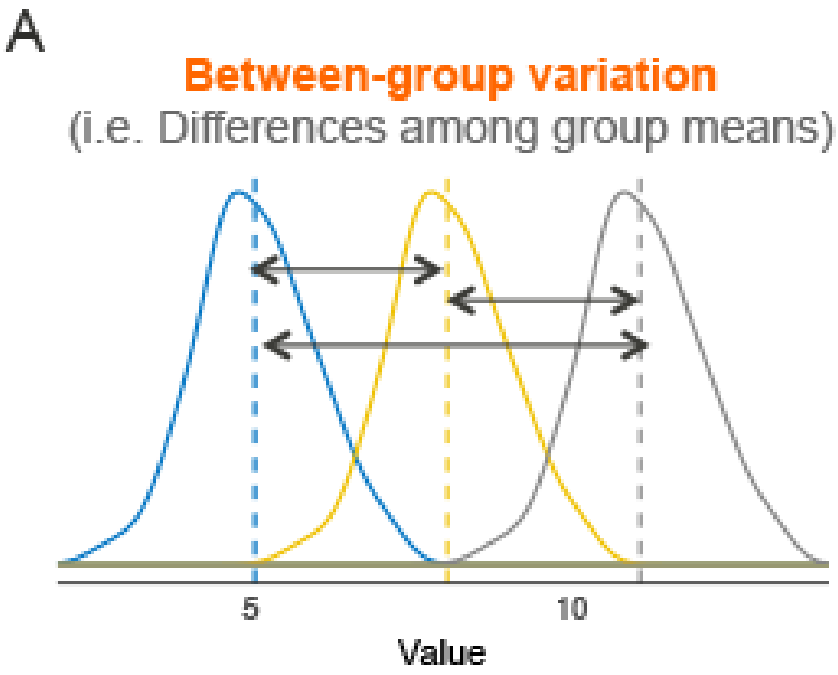
- Analysis of variance (ANOVA) is a statistical technique that is used to check if two or more groups are significantly different from each other.

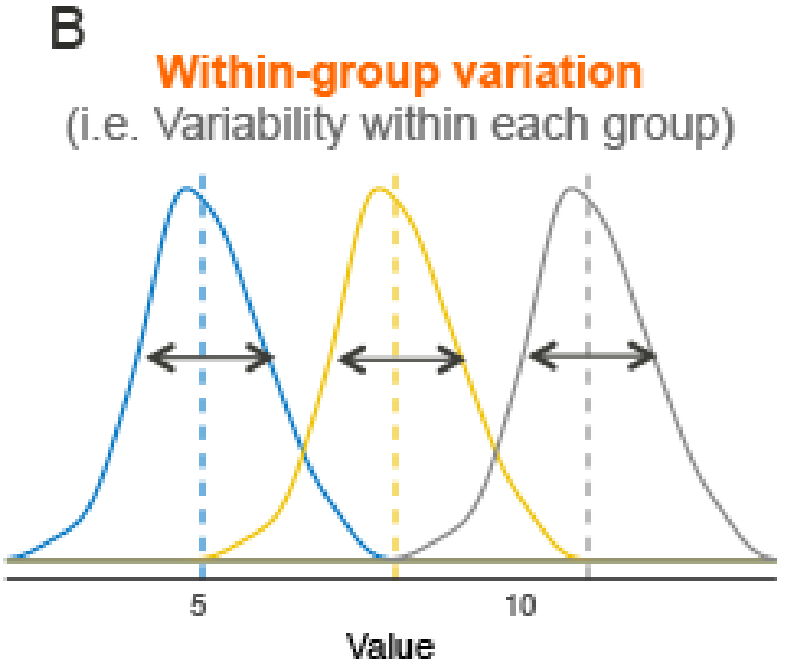
- ANOVA checks the impact of one or more factors by comparing the variances of different samples



Variability

- ANOVA considers between group variation and within group variation.
- Total variation = between-group variation + within-group variation.





ANOVA Hypothesis

- Like t-test and chi-square test, ANOVA also uses a Null hypothesis and an Alternate hypothesis.

- The Null hypothesis in ANOVA is valid when all the sample means are equal, or they don't have any significant difference. Thus, they can be considered as a part of a larger set of the population.

- On the other hand, the alternate hypothesis is valid when at least one of the sample means is different from the rest of the sample means. In mathematical form, they can be represented as:

$$H_o : \mu_1 = \mu_2 = \cdots = \mu_L \qquad \textbf{Null hypothesis}$$

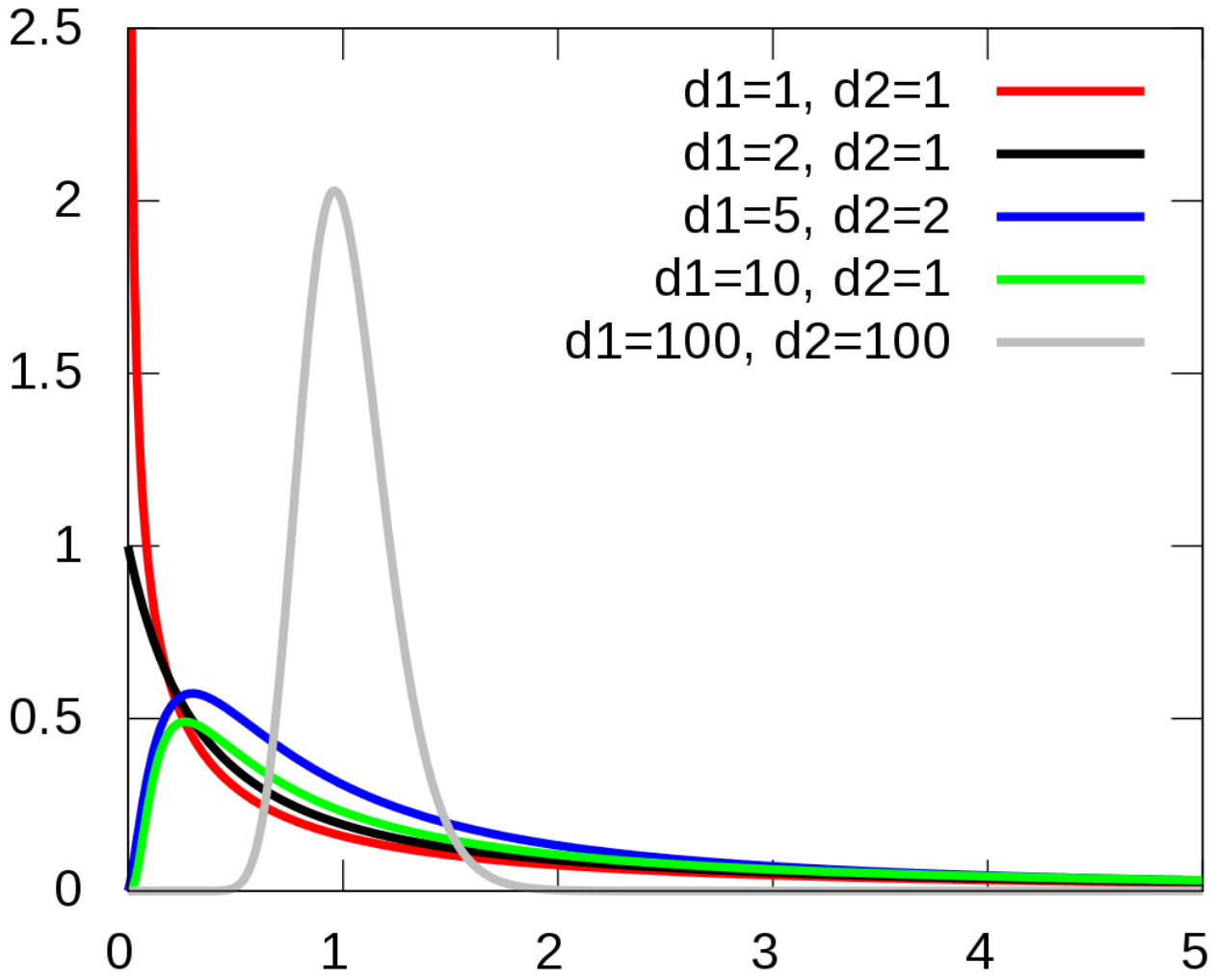
$$H_1 : \mu_l \neq \mu_m \qquad \textbf{Alternate hypothesis}$$

F Statistics

- The statistic which measures if the means of different samples are significantly different or not is called the F-Ratio.
- Lower the F-Ratio, more similar are the sample means. In that case, we cannot reject the null hypothesis.
- **F = Between group variability / Within group variability**
- The numerator term in the F-statistic calculation defines the between-group variability. As we read earlier, as between group variability increases, sample means grow further apart from each other. In other words, the samples are more probable to be belonging to totally different populations.

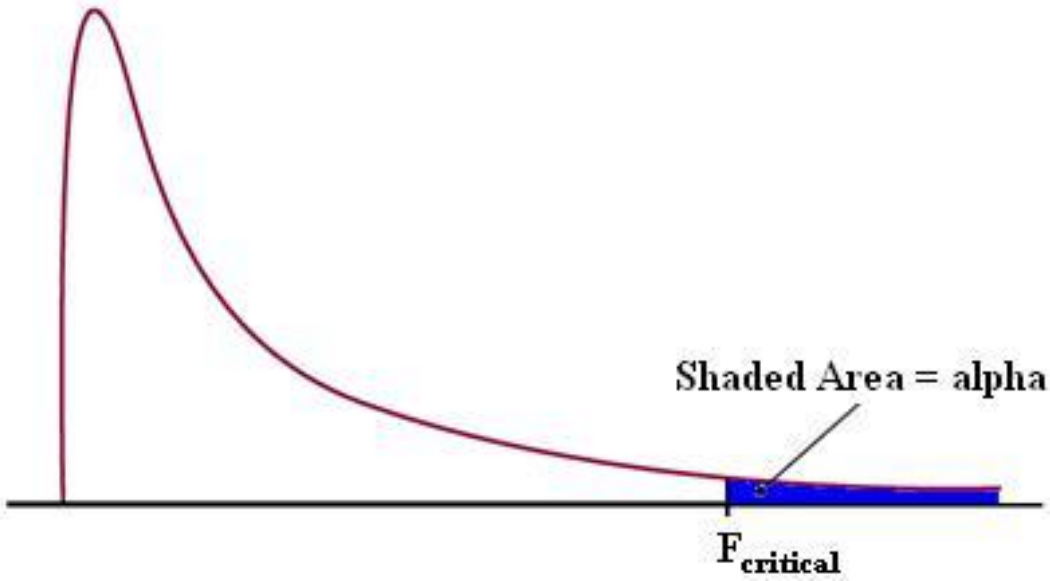
F Distribution

- The F distribution is the probability distribution associated with the f statistic.

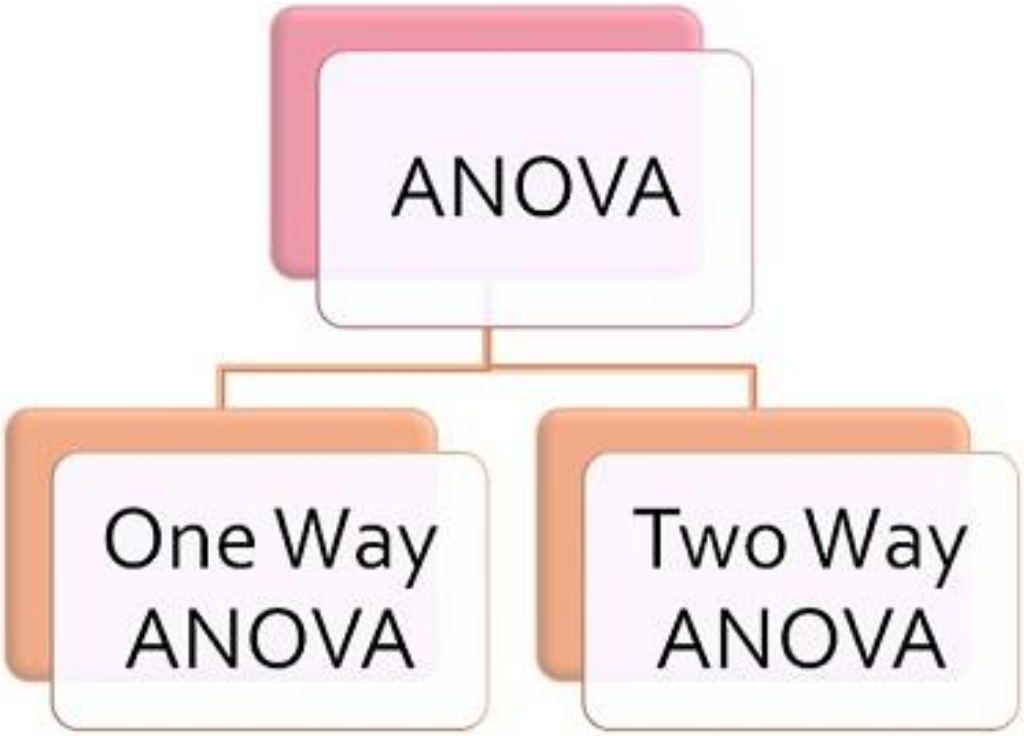


Determine P Value

- This F-statistic calculated here is compared with the F-critical value for making a conclusion.
- If the value of the calculated F-statistic is more than the F-critical value, then we reject the null hypothesis.
- Alternatively, one can compute the p-value of F statistics. If the p value is less than the significance level, then we reject the null hypothesis.
- You can compute the critical F value from this online tool
- https://www.danielsoper.com/statcalc/calculator.aspx?id=4
- https://www.danielsoper.com/statcalc/calculator.aspx?id=7



ANOVA (*)



One-Way ANOVA (*)

Assumptions:

- Normally Distributed
- Independent Observations
- Equivalent Variance

One Way ANOVA

- The one-way ANOVA is used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups whilst considering only one independent variable or factor.
- Example of data for one-way ANOVA could be:

Detergent A	Detergent B	Detergent C
15	18	10
12	14	9
10	18	7
6	12	5

Activity: One Way ANOVA

- Test the hypothesis of whether there is any significant differences among the 3 shampoos below using one-way ANOVA test.

Shampoo A	Shampoo B	Shampoo C
36.6	17.5	15.0
39.2	20.6	10.4
30.4	18.7	18.9
37.1	25.7	10.5
34.1	22.0	15.2

- You can perform the one-way ANOVA using the tool below https://www.socscistatistics.com/tests/anova/default2.aspx

How Variances is Calculated?

Solution: $\bar{x}_{1.} = \frac{36.6+39.2+30.4+37.1+34.1}{5} = 35.48$, $\bar{x}_{2.} = \frac{17.5+20.6+18.7+25.7+22.0}{5} = 20.9$, $\bar{x}_{3.} = \frac{15.0+10.4+18.9+10.5+1.2}{5} = 14$ and $\bar{x}_{..} = \frac{35.48+20.9+14}{3} = 23.46$.

$$\text{SS(total)} = \sum_{i=1}^{3}\sum_{j=1}^{5}(x_{ij} - \bar{x}_{..})^2 = (36.6 - 23.46)^2 + (39.2 - 23.46)^2 + \ldots$$
$$+ (10.5 - 23.46)^2 + (15.2 - 23.46)^2 = 1340.456$$

$$\text{SS(within)} = \sum_{i=1}^{3}\sum_{j=1}^{5}(x_{ij} - \bar{x}_{i.})^2 = (36.6 - 35.48)^2 + \ldots + (34.1 - 35.48)^2$$
$$+ (17.5 - 20.9)^2 + \ldots + (22.0 - 20.9)^2 + (15.0 - 14)^2 + \ldots + (15.2 - 14)^2$$
$$= 137.828$$

$$\text{SS(between)} = \sum_{i=1}^{3} 5 \times (\bar{x}_{i.} - \bar{x}_{..})^2 = 5 \times ((35.48 - 23.46)^2 + (20.9 - 23.46)^2 + (14 - 23.46)^2$$
$$= 1202.628$$

SS: Sum of Squares = Variance
SS(total) = SS(within)+SS(between)

ANOVA Table

ANOVA table:

Source	Sum of Squares	Degree of Freedom	Mean Square	F value
Between	1202.628	2	601.314	52.35
Within	137.828	12	11.486	
Total	1340.456	14		

MS: Mean Squares = Sum of Squares/DF
F = MS(Between)/MS(Within)

ANOVA Table

Summary ANOVA

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Between	SS_B	$K-1$	$MS_B = \frac{SS_B}{K-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-K$	$MS_W = \frac{SS_W}{N-K}$	
Total	$SS_T = SS_B + SS_W$	$N-1$		

Example: Lamb Weight Gain (*)

The weight gain of lambs on three different diets over a 2-week period.

Weight Gain (lbs.)		
Diet 1	Diet 2	Diet 3
8	9	15
16	16	10
9	21	17
	11	6
	18	

Question: any difference between these three different diets?

Lamb Weight Gain Example (*)

Source	Sum of Squares	Degree of Freedom	Mean Square	F value
Between	36	2	18	0.7714
Within	210	9	70/3	
Total	246	11		

Example: Lamb Weight Gain (*)

$$F = \frac{36/(3-1)}{210/(12-3)} = 0.7714$$

P-value=P(F>0.7714)

=1-F.DIST(0.7714,2,9,TRUE)=0.49>0.05

We do not reject null hypothesis. (No significant difference between any of the groups)

Exercise 1 (*)

A traffic engineering study on traffic delay was conducted at intersections with signals on urban streets. Three types of traffic signals were utilized in the study: (1) pretimed, (2) semi-actuated, and (3) fully actuated. Five intersections were used for each type of signal. The measure of traffic delay used in the study was the average stopped time per vehicle at each of the intersections (seconds/vehicle).

Exercise 1 (*)

Pretimed	Semi-actuated	Fully actuated
36.6	17.5	15.0
39.2	20.6	10.4
30.4	18.7	18.9
37.1	25.7	10.5
34.1	22.0	15.2

Compute the ANOVA table and test the hypothesis of no difference among the mean traffic delays of the signal types with the suitable test at the 0.05 level of significance.

Module5a – 1way

Two Way ANOVA

- The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors).
- The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.
- For example, you could use a two-way ANOVA to understand whether there is an interaction between gender and drug level on anxiety amongst patients, where gender (males/females) and drug level (1,2,3)

Patents	Drug 1	Drug 2	Drug 3
Male	8	10	8
	4	8	6
	0	6	4
Female	14	4	15
	10	2	12
	6	0	9

Exercise 2 (*)

A school district uses four different methods of teaching their students how to read and wants to find out if there is any significant difference between the reading scores achieved using the four methods. It creates a sample of 8 students for each of the four methods. The reading scores achieved by the participants in each group are as follows:

Exercise 2 (*)

Method 1	Method 2	Method 3	Method 4
51	82	79	85
87	91	84	80
50	92	74	65
48	80	98	71
79	52	63	67
61	79	83	51
53	73	85	63
54	74	58	93

Use Excel to compute the ANOVA table

Module5b - reading

Hypotheses for Two Way ANOVA

- Because the two-way ANOVA consider the effect of two categorical factors, and the effect of the categorical factors on each other, there are three pairs of null or alternative hypotheses for the two-way ANOVA. For example

- H0: The means of all drug levels are equal
- H1: The mean of at least drug level is different

- H0: The means of the gender groups are equal
- H1: The means of the gender groups are different

- H0: There is no interaction between the drug level and gender
- H1: There is interaction between the drug level and gender

Activity: Two Way ANOVA

- A physiologist was interested in learning whether smoking history and different types of stress tests influence the timing of a subject's maximum oxygen uptake, as measured in minutes.

- The researcher classified a subject's smoking history as either heavy smoking, moderate smoking, or non-smoking. He was interested in seeing the effects of three different types of stress tests a test performed on a bicycle, a test on a treadmill, and a test on steps.

- The physiologist recruited 9 non-smokers, 9 moderate smokers, and 9 heavy smokers to participate in his experiment, for a total of n = 27 subjects.

- He then randomly assigned each of his recruited subjects to undergo one of the three types of stress test

Activity: Two Way ANOVA

- Here are his resulting data:

Sample History	Bicycle	Treadmill	Step Test
Non Smoker	12.8	16.2	22.6
	13.5	18.1	19.3
	11.2	17.8	18.9
Moderate Smoker	10.9	15.5	20.1
	11.1	13.8	21
	9.8	16.2	15.9
Heavy Smoker	8.7	14.7	16.2
	9.2	13.2	16.1
	7.5	8.1	17.8

- You can use the following online tool for two-way ANOVA to analyze the above data
  http://vassarstats.net/anova2u.html

One Way vs Two Way ANOVA

- A one-way ANOVA is primarily designed to enable the equality testing between three or more means. A two-way ANOVA is designed to assess the interrelationship of two independent variables on a dependent variable.

- A one-way ANOVA only involves one factor or independent variable, whereas there are two independent variables in a two-way ANOVA.

- In a one-way ANOVA, the one factor or independent variable analyzed has three or more categorical groups. A two-way ANOVA instead compares multiple groups of two factors.

- One-way ANOVA need to satisfy only two principles of design of experiments, i.e. replication and randomization. As opposed to Two-way ANOVA, which meets all three principles of design of experiments which are replication, randomization, and local control (all extraneous sources of variation are brought under control)

Two-way ANOVA with vs without Replication (*)

- Two-way ANOVA without replication is where only the two main 'treatments' or effects can be tested. There is only 1 value for each combination between the levels of the 2 factors.

- Two-way ANOVA with replication, we can measure and characterize the two main effects as well as the interaction between them. There is more than 1 value for each combination between the levels of the 2 factors.

Two-way ANOVA without Replication (*)

A new fertilizer has been developed to increase the yield on crops, and the makers of the fertilizer want to better understand which of the three formulations (blends) of this fertilizer are most effective for wheat, corn, soybeans and rice (crops). They test each of the three blends on one sample of each of the four types of crops. The crop yields for the 12 combinations.

Two-way ANOVA without Replication (*)

	Wheat	Corn	Soy	Rice
Blend X	123	138	110	151
Blend Y	145	165	140	167
Blend Z	156	176	185	175

Two Factors: Blends & Crops
Blends: 3 levels
Crops: 4 levels

No Replication: only one sample for each combination

Module5c – 2_way_ANOVA

Two-way ANOVA without Replication (*)

Hypotheses:

H**0**: there is no significant difference in yield between the (population) means of the blends

OR

H**0**: there is no significant difference in yield between the (population) means for the crop types

Test about Rows (*)

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	3629.16667	2	1814.58333	12.8264284	0.00681106	5.14325285
Columns	1116.91667	3	372.305556	2.63165129	0.14456122	4.75706266
Error	848.833333	6	141.472222			
Total	5594.91667	11				

Since the p-value for the rows = 0.0068 < 0.05 we reject the null hypothesis, and so at the 95% level of confidence we conclude there is significant difference in the yields produced by the three blends.

Test about Columns (*)

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	3629.16667	2	1814.58333	12.8264284	0.00681106	5.14325285
Columns	1116.91667	3	372.305556	2.63165129	0.14456122	4.75706266
Error	848.833333	6	141.472222			
Total	5594.91667	11				

Since the p-value for the columns = 0.1446 > 0.05, we can't reject the null hypothesis, and so at 95% level of confidence we conclude there is no significant difference in the yields for the four crops studied.

Exercise (*)

An experiment was conducted to evaluate the effect of different detergents and water temperatures on the cleanliness of ceramic substrates. The experimenter selected three different detergents based on their pH levels and conducted a series of experiments at four different water temperatures.

Exercise (*)

Cleanliness was quantified by measuring the contamination of a distilled water beaker after rinsing the parts cleaned using each treatment combination. Using Excel to find the ANOVA table.

	DETERGENT A	DETERGENT B	DETERGENT C
Cold	15	18	10
Cool	12	14	9
Warm	10	18	7
Hot	6	12	5

Module5d - Cleaning

Two-way ANOVA with Replication (*)

Repeat the analysis of Two-way ANOVA with Replication, but this time, each combination of blend and crop has a sample of size 5.

Two-way ANOVA with Replication (*)

	Crop			
Fertilizer	Wheat	Corn	Soy	Rice
BlendX	123	128	166	151
	156	150	178	125
	112	174	187	117
	100	116	153	155
	168	109	195	158
BlendY	135	175	140	167
	130	132	145	183
	176	120	159	142
	120	187	131	167
	155	184	126	168
BlendZ	156	186	185	175
	180	138	206	173
	147	178	188	154
	146	176	165	191
	193	190	188	169

Module5e – 2_way_Replication

Two-way ANOVA with Replication (*)

a) Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that any differences between the effectiveness of the fertilizer for the different blends?

b) Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that any differences between the effectiveness of the fertilizer for the different crops?

c) Is there evidence of an interaction between crop and blend?

Two-way ANOVA with Replication (*)

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	8782.9	2	4391.45	9.93334716	0.00024549	3.19072734
Columns	3411.65	3	1137.21667	2.5723549	0.06494382	2.79806064
Interaction	6225.9	6	1037.65	2.34713766	0.04555549	2.29460131
Within	21220.4	48	442.091667			
Total	39640.85	59				

a) Since the p-value (blends) = 0.00025<0.05, we reject the null hypothesis, and conclude that the blends are statistically different.

Two-way ANOVA with Replication (*)

b) Since the p-value (crops) = 0.0649 > 0.05 we can't reject the Factor A null hypothesis, and so conclude (with 95% confidence) that there are no significant differences between the effectiveness of the fertilizer for the different crops.

c) We also see that the p-value (interactions) =0.0456 <0.05, and so conclude there are significant differences in the interaction between crop and blend.

Exercise (*)

A physiologist was interested in learning whether smoking history and different types of stress tests influence the timing of a subject's maximum oxygen uptake, as measured in minutes. The researcher classified a subject's smoking history as either heavy smoking, moderate smoking, or non-smoking.

Exercise (*)

He was interested in seeing the effects of three different types of stress tests — a test performed on a bicycle, a test on a treadmill, and a test on steps. The physiologist recruited 9 non-smokers, 9 moderate smokers, and 9 heavy smokers to participate in his experiment, for a total of $n = 27$ subjects. He then randomly assigned each of his recruited subjects to undergo one of the three types of stress test.

Exercise (*)

	Test		
Smoking History	Bicycle (1)	Treadmill (2)	Step Test (3)
Nonsmoker (1)	12.8	16.2	22.6
	13.5	18.1	19.3
	11.2	17.8	18.9
Moderate (2)	10.9	15.5	20.1
	11.1	13.8	21
	9.8	16.2	15.9
Heavy (3)	8.7	14.7	16.2
	9.2	13.2	16.1
	7.5	8.1	17.8

Module5f - smoking

Exercise (*)

a) Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that smoking history influences the time to maximum oxygen uptake?

b) Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that the type of stress test influences the time to maximum oxygen uptake?

c) Is there evidence of an interaction between smoking history and the type of stress test?
Time: 10 mins

Topic 6
Regression

Regression Analysis (*)

We may want to use a person's height, gender, race, etc. to predict the person's weight.

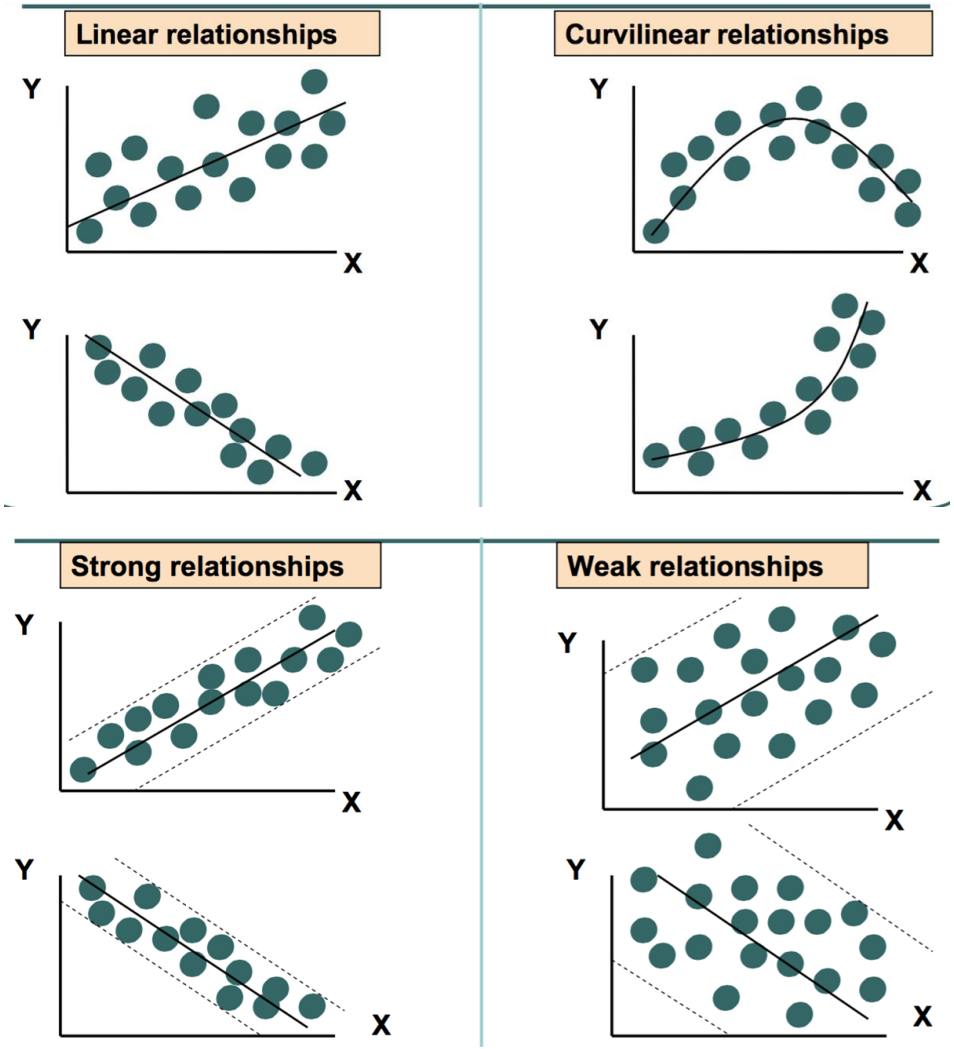
- One variable, denoted *x*, is regarded as the predictor or independent variable.
- The other variable, denoted *y*, is regarded as the response, outcome, or dependent variable.

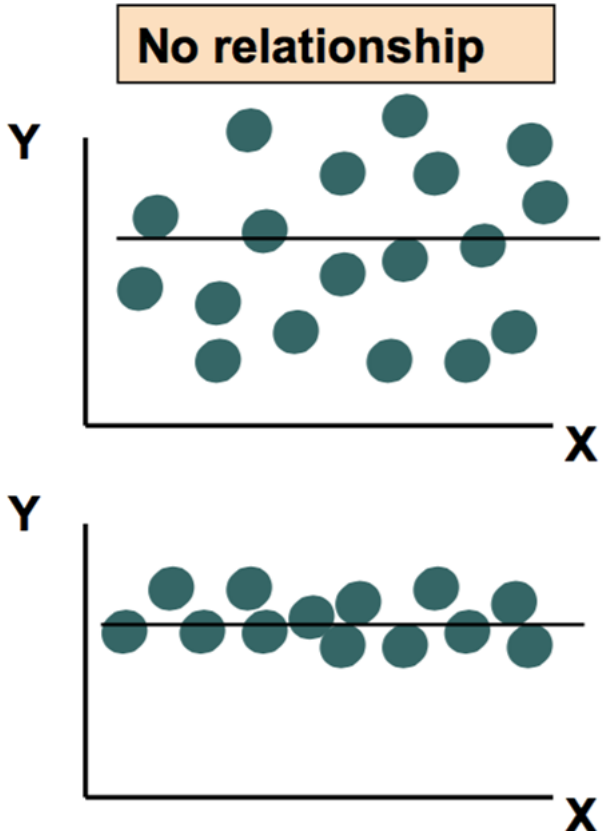
e.g. y=weight x=height

Regression Analysis (*)

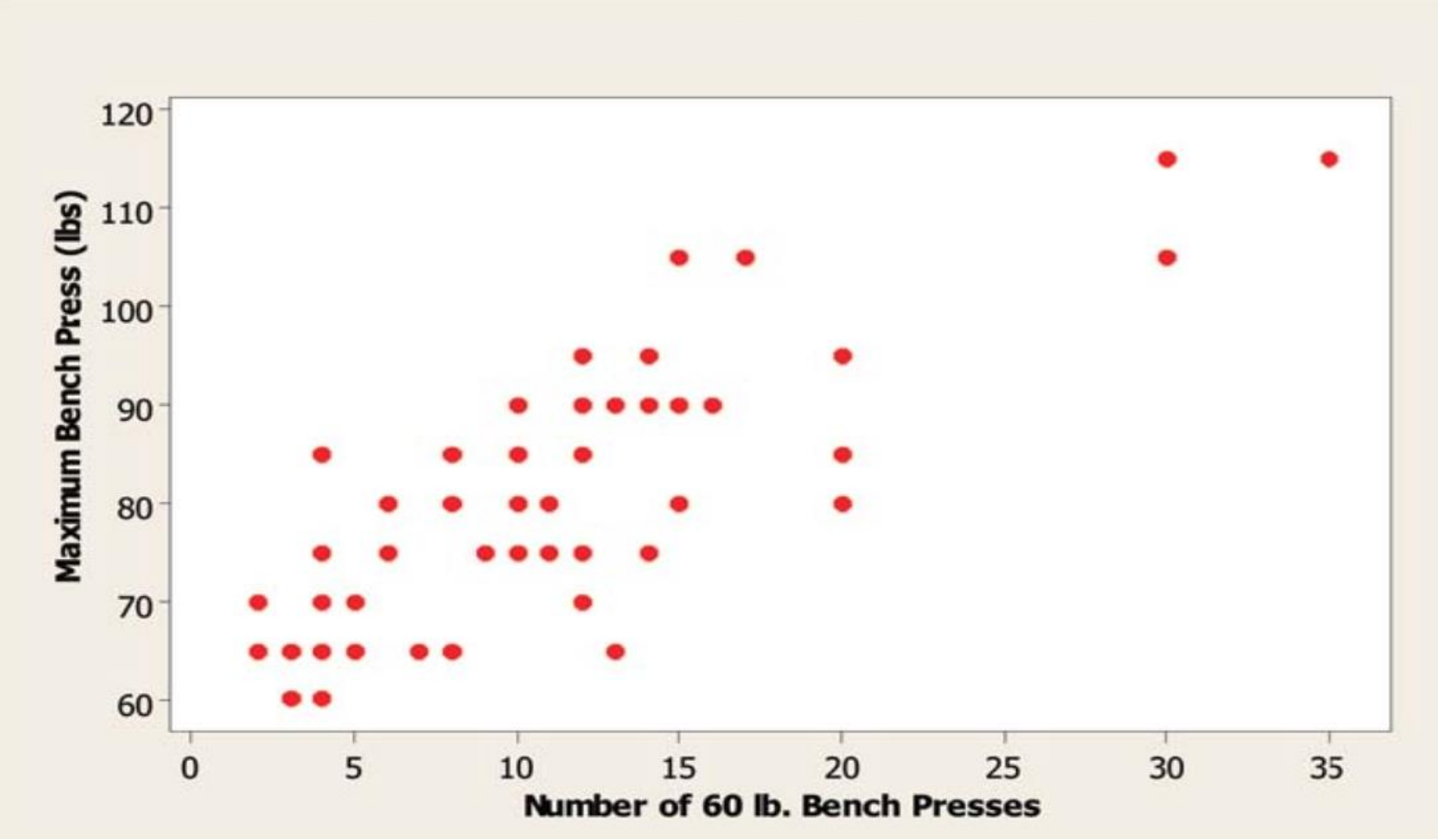
- Visualize the data on a scatterplot.
- A *scatterplot* is a graphical display of the relationship between two variables.
- Regression line is the ‘best-fit’ line through a scatterplot.

Types of Relationships (*)

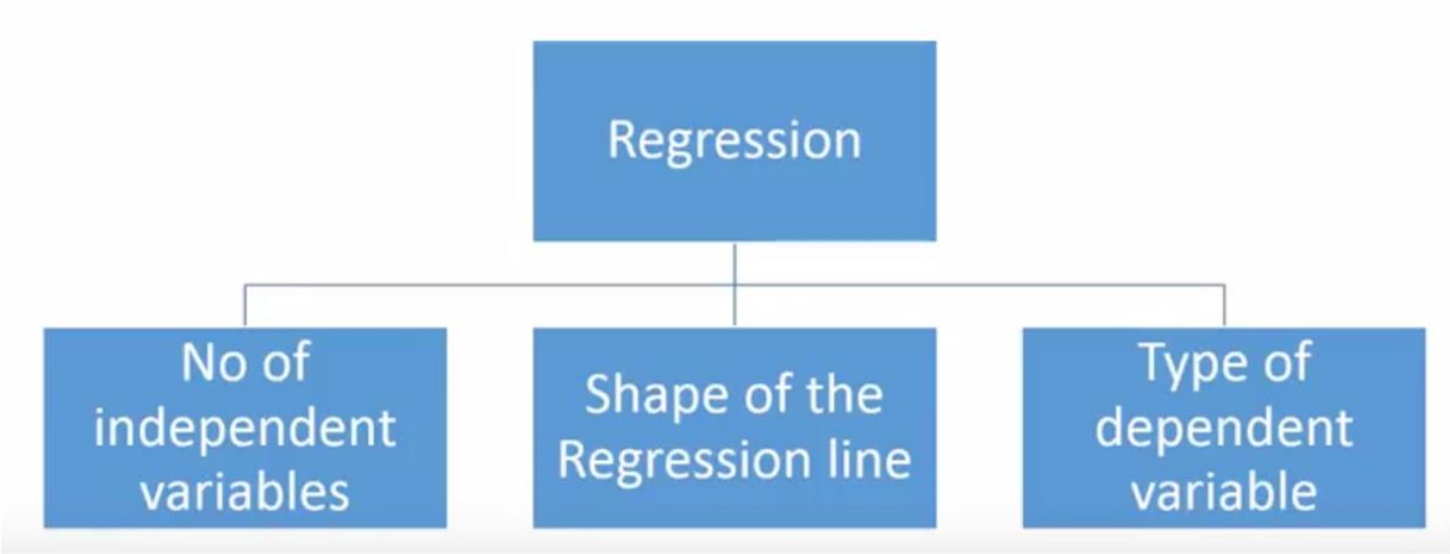




Example (*)

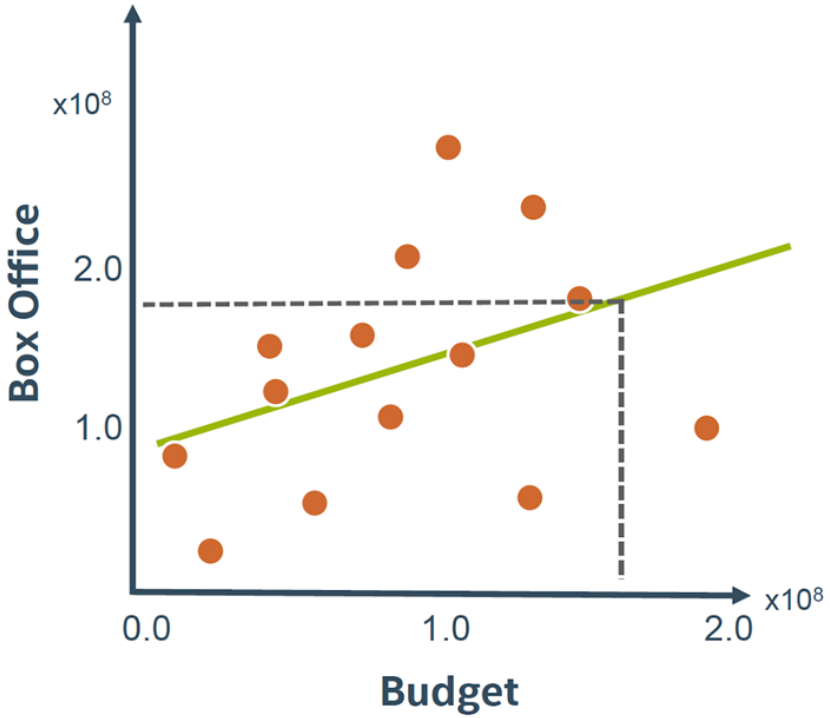

Maximum Bench Press (lbs)
120
110
100
90
80
70
60
0
5
10
15
20
25
30
35
Number of 60 lb. Bench Presses

Types of Regression (*)



Linear Regression

- Linear regression is the most common regression model. Many predictive models use linear regression models
- You can use a linear regression model to predict the box office from the budget.



$$y_\beta(x) = \beta_0 + \beta_1 x$$

β_0= 80 million, β_1= 0.6

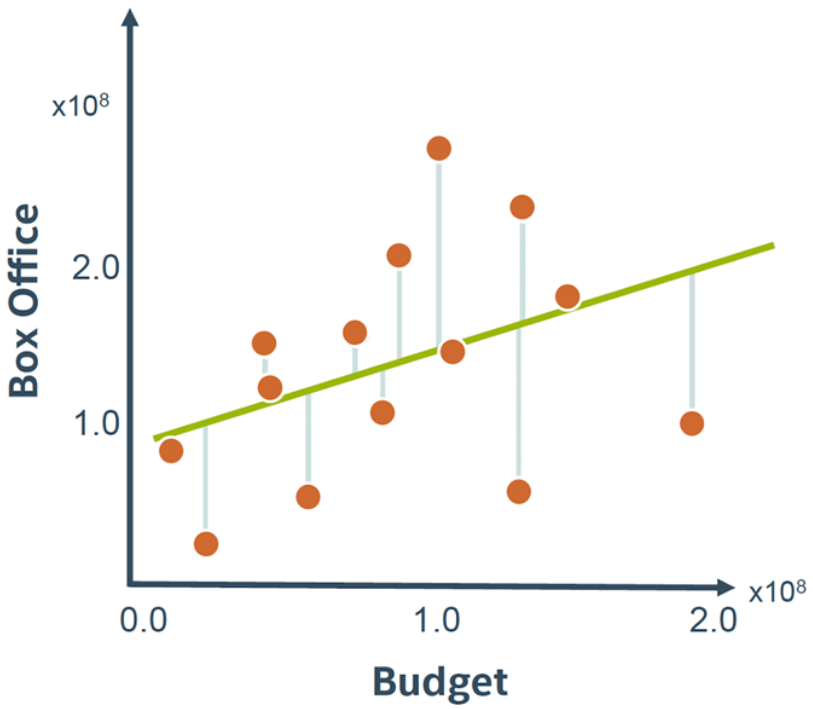
Predict 175 Million Gross for 160 Million Budget

Interpretation of the coefficients within the regression equation

- a :the *y*-intercept of the regression line.
- b : the slope of the regression line.
- b: represents the estimated increase in *y* per unit increase in *x*. Note that the increase may be negative which is reflected when b is negative.

Residues

- Residue is the difference between the predicted value and actual value



$$y_\beta\left(x_{obs}^{(i)}\right) - y_{obs}^{(i)}$$

Predicted value

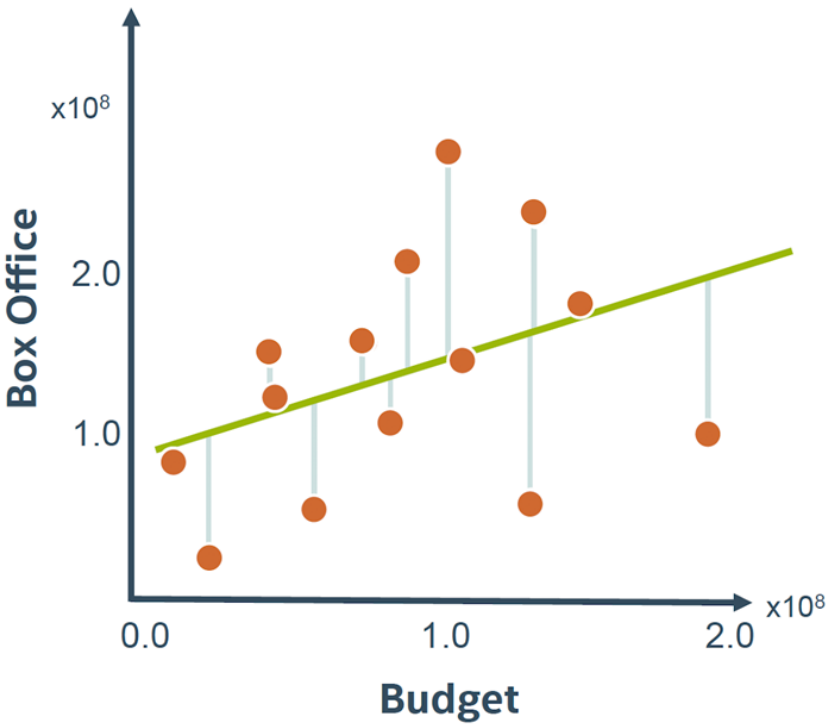
Observed value

$$\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}$$

Mean Square Error

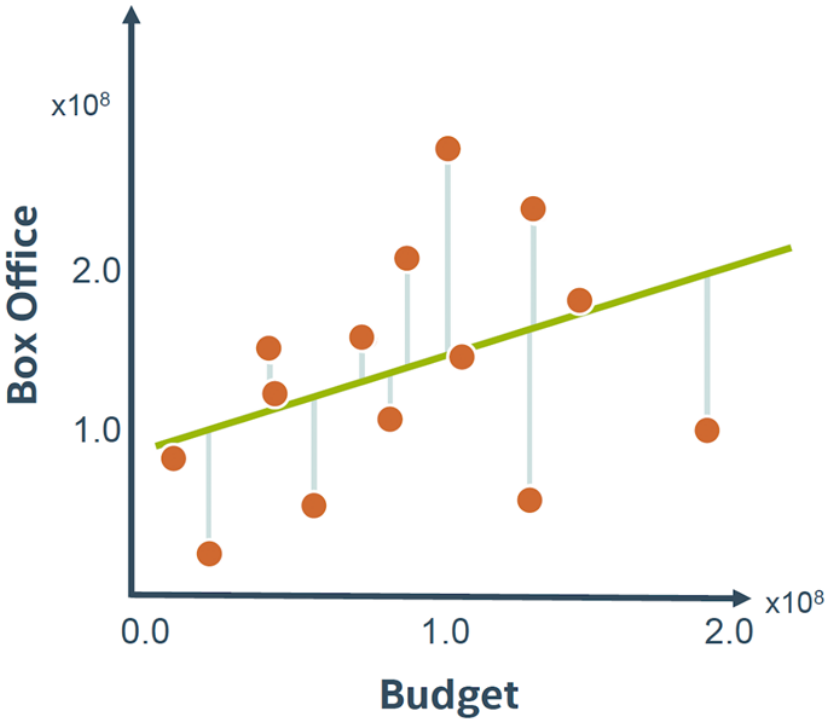
- Mean Square Error (MSE) is the common loss function to measure how good is the linear regression model.

$$\frac{1}{m}\sum_{i=1}^{m}\left(\left(\beta_0 + \beta_1 x_{obs}^{(i)}\right) - y_{obs}^{(i)}\right)^2$$



Minimum Mean Square Error

- Regression aims to minimize the MSE to find the best linear regression model.

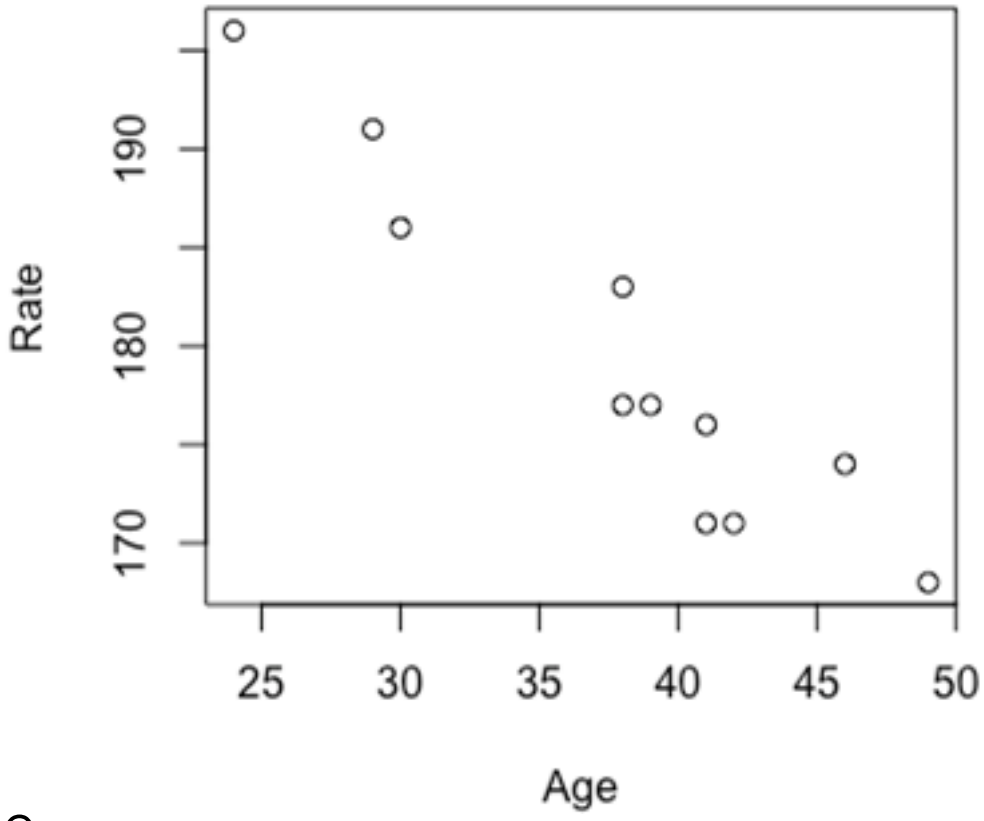


$$\min_{\beta_0,\beta_1} \frac{1}{m} \sum_{i=1}^{m} \left(\left(\beta_0 + \beta_1 x_{obs}^{(i)} \right) - y_{obs}^{(i)} \right)^2$$

Example (*)

The following data are 11 randomly selected people's maximum heart rate (y) and their age (x)

Age	Rate
30	186
38	183
41	171
38	177
29	191
39	177
46	174
41	176
42	171
24	196
49	168



Module6a - Heartrate

Example (*)

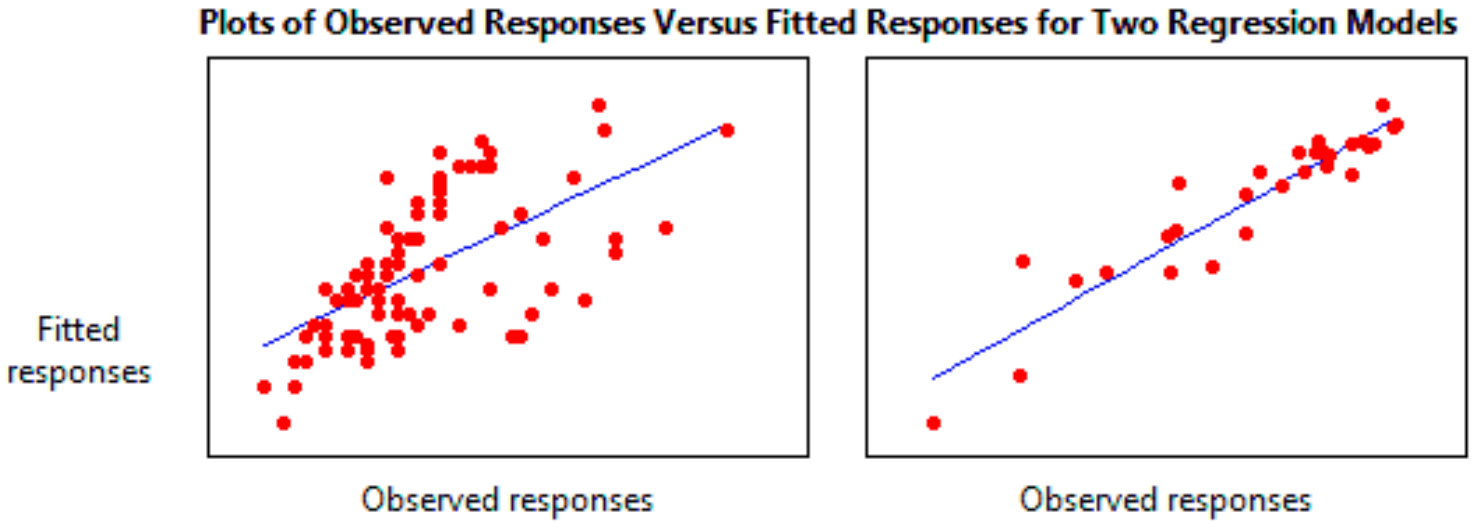
The regression equation is:

$$rate = 221.66 - 1.123 \times age$$

The coefficient of x is the coefficient of age, which is -1.123. It means for every one-year increase in age, the estimated increase in maximum heart rate is: -1.123 (equivalently, the estimated decrease in maximum heart rate is 1.123.)

R Square (Goodness Of Fit)

- R-squared is a statistical measure of how close the data are to the fitted regression line.
- R-squared = Explained variation / Total variation
- R-squared is always between 0 and 1
  - 0 indicates that the model explains none of the variability of the response data around its mean.
  - 1 indicates that the model explains all the variability of the response data around its mean.



Activity: Regression

- The data are collected at the end of an introductory statistics course. The table shows the data for the eight males in the class on these variables and on the number of class lectures for the course that the student reported skipping during the term.
- Investigate the relationship between x=study time and y=GPA. Find the prediction equation and interpret the slope.
- You can use the following one tool https://keisan.casio.com/exec/system/14059929550941

Student	Study Time	Grade Point
1	14	2.8
2	25	3.6
3	15	3.4
4	5	3.0
5	10	3.1
6	12	3.3
7	5	2.7
8	21	3.8

Exercise (*)

Student	Study Time	GPA	Skipped
1	14	2.8	9
2	25	3.6	0
3	15	3.4	2
4	5	3.0	5
5	10	3.1	3
6	12	3.3	2
7	5	2.7	12
8	21	3.8	1

Exercise (*)

a) Investigate the relationship between x=study time and y=GPA. Find the prediction equation and interpret the slope.

b) Find the predicted GPA for a student who studies 25 hours per week using equation in a).

c) Find and interpret the residual for Student 2, who reported x=25 using equation in a).

Exercise (*)

d) Investigate the relationship between x=number of classes skipped and y=GPA. Find the prediction equation and interpret the slope.

e) Find the predicted GPA and residual for Student 1 equation in d).

Module6b - GPA

Topic 7
Correlation Analysis

What is Covariance?

- Variance is a measure of the variability or spread in a set of data
- We use the following formula to compute variance for population and sample, respectively.

$$Var(x) = \frac{\sum(x - \overline{x})^2}{N} \qquad Var(x) = \frac{\sum(x - \overline{x})^2}{N - 1}$$

- **Covariance** is a measure of the extent to which corresponding elements from two sets of ordered data move in the same direction.
- We use the following formula to compute covariance for population and sample respectively

$$Cov(x, y) = \frac{\sum(x - \overline{x})(y - \overline{y})}{N} \qquad Cov(x, y) = \frac{\sum(x - \overline{x})(y - \overline{y})}{N - 1}$$

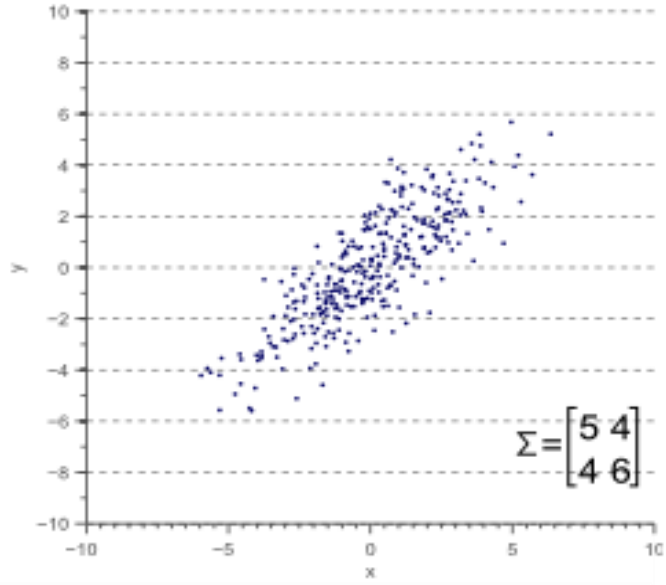
Covariance Matrix

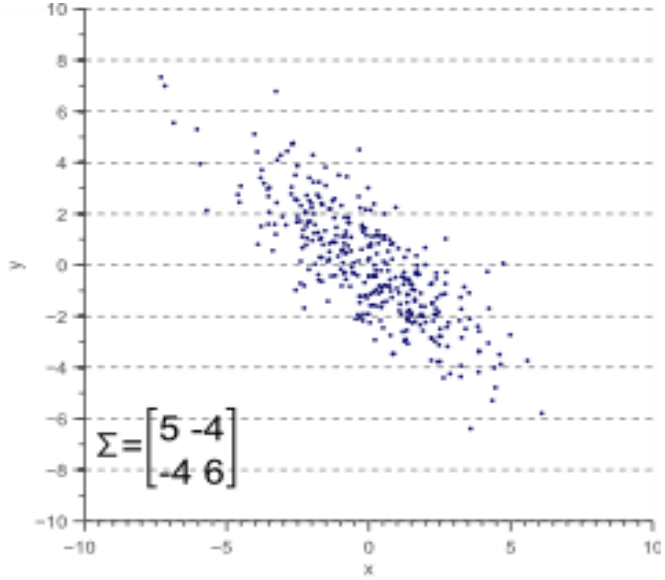
- Variance and covariance are often displayed together in a covariance matrix given as follows:

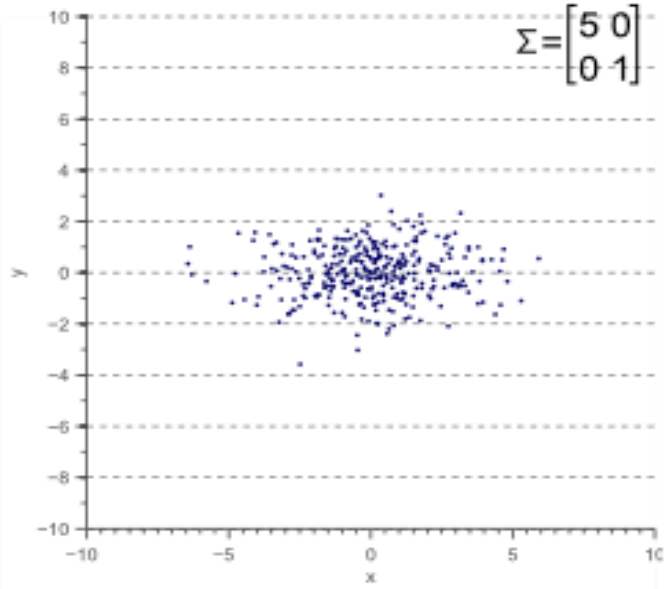
$$\text{Cov}(A) = \begin{bmatrix} \frac{\sum (x_i - \bar{X})(x_i - \bar{X})}{N} & \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{N} \\ \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{N} & \frac{\sum (y_i - \bar{Y})(y_i - \bar{Y})}{N} \end{bmatrix}$$

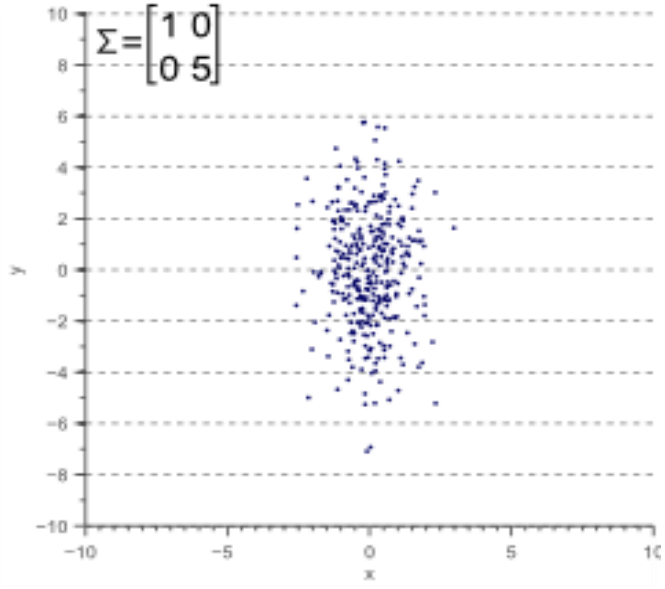
$$= \begin{bmatrix} \text{Cov}(X,X) & \text{Cov}(Y,X) \\ \text{Cov}(X,Y) & \text{Cov}(Y,Y) \end{bmatrix}$$

Covariance Matrix Visualization









Interpreting Covariance Value (*)

- Zero : the data sets don't vary together
- Positive: the data sets tend to move together
- Negative: the data sets tend to move in opposite directions

Example (*)

Suppose you take a sample of stock returns from the Excelsior Corporation and the Adirondack Corporation from the years 2008 to 2012

Year	Excelsior Corp. Annual Return (percent) (*X*)	Adirondack Corp. Annual Return (percent) (*Y*)
2008	1	3
2009	–2	2
2010	3	4
2011	0	6
2012	3	0

Module7a – Covariance Example

Activity: Covariance

Compute the covariance for the following data

X: 90,90,60,60,30
Y: 60,90,60,60,30

You can use the online covariance calculator

https://planetcalc.com/8125/

Exercise (*)

Find the covariance matrix in the blood pressure example

Module7b - bloodpressure

What is Correlation?

- The correlation coefficient is also known as the Pearson product-moment correlation coefficient, or Pearson's correlation coefficient.

- It is obtained by dividing the covariance of the two variables by the product of their standard deviations.

$$Corr(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

- The values of the correlation coefficient can range from -1 to +1. The closer it is to +1 or -1, the more closely are the two variables are related.

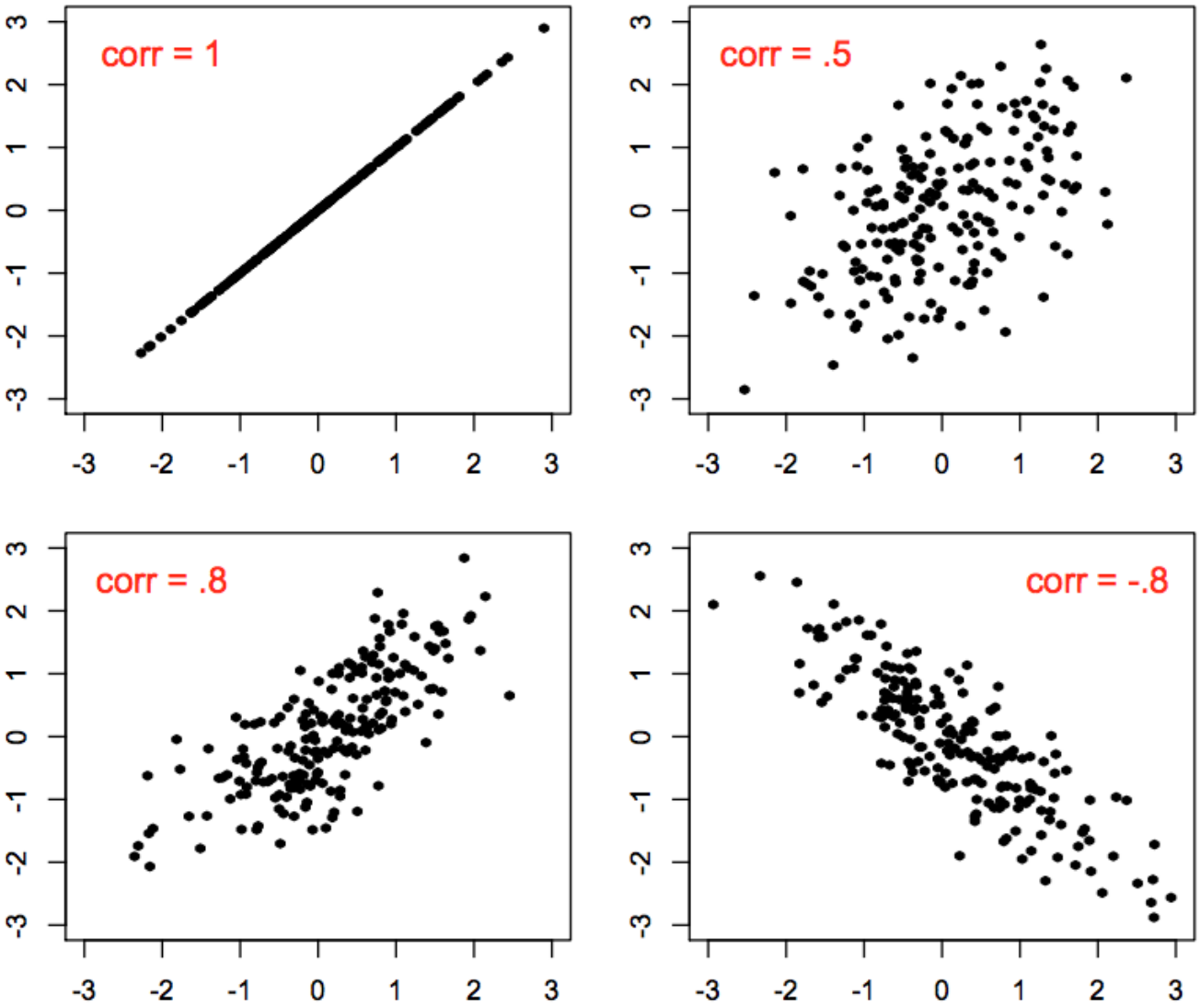
- The positive sign signifies the direction of the correlation i.e. if one of the variables increases, the other variable is also supposed to increase.

Correlation Matrix

- For multiple variables, we can display all the correlation coefficients in the matrix form as below:

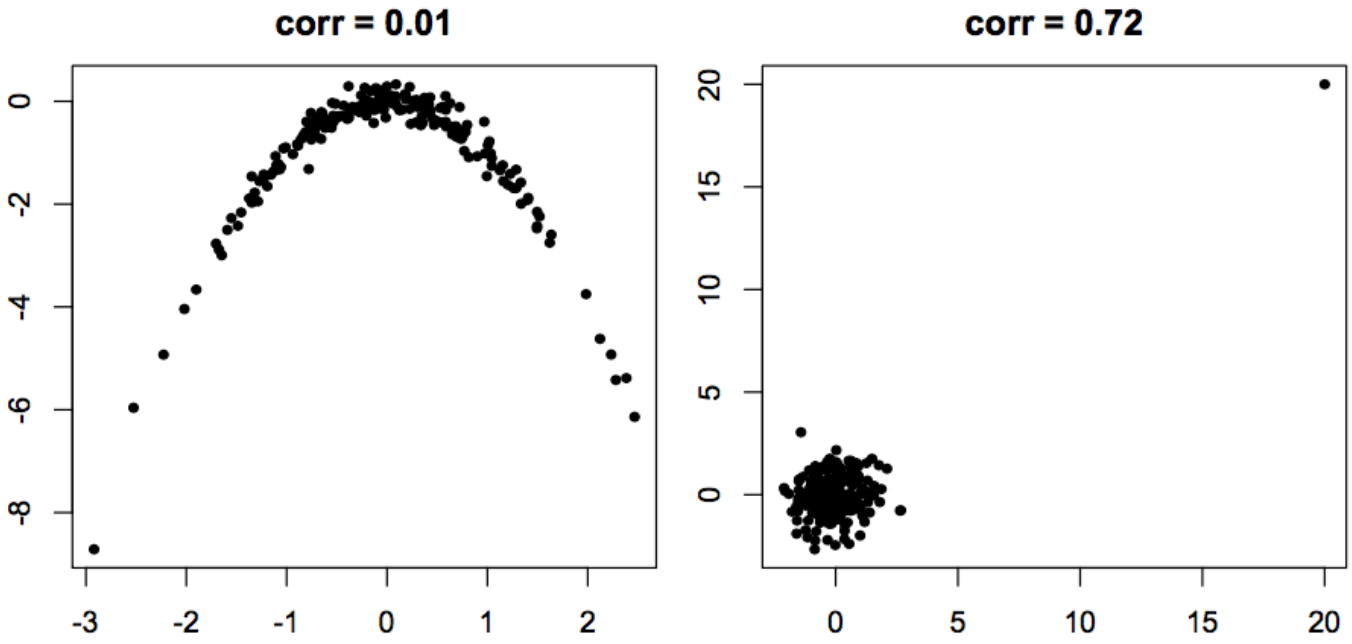
$$\begin{bmatrix} 1 & \text{Corr}(X,Y) & \text{Corr}(X,Z) \\ \text{Corr}(X,Y) & 1 & \text{Corr}(Y,Z) \\ \text{Corr}(X,Z) & \text{Corr}(Y,Z) & 1 \end{bmatrix}$$

Correlation Coefficient



Correlation Coefficient r (*)

Only measures linear relationships:
r= 0 does not mean the variables are not related!



Interpreting Correlation Value (*)

Strength of linear relationship	Positive	Negative
Very strong	0.8 to 1	−0.8 to −1
Strong	0.4 to 0.79	−0.4 to −0.79
Weak	0.2 to 0.39	−0.2 to −0.39
Little / no relationship	0 to 1.9	0 to −0.19

Activity: Correlation

Compute the Pearson correlation coefficient for the following data

X: 90,90,60,60,30

Y: 60,90,60,60,30

You can use the online correlation calculator

https://www.socscistatistics.com/tests/pearson/default2.aspx

Exercise (*)

Find the correlation matrix in the blood pressure example

Module7b - bloodpressure

Summary Q&A

Course Feedback

https://goo.gl/R2eumq

Thank You!

Marcus Lee Yi Qing
87119800
makasulee@gmail.com