# Intro to Probability and Statistics

**Unit 6**

*Working draft*

**Nepal Data Literacy Program, 2019**

Organized by

THE WORLD BANK
IBRD • IDA

Supported by

UKaid
from the British people

**Unit 6: Module 1: Basic Probability**

**Unit 6: Module 2: Binomial, Normal Distribution, Statistical Inference**

**Unit 6: Module 3: Univariate Analysis - Confidence Interval**

**Unit 6: Module 4: Univariate Analysis - Hypothesis Testing**

**Unit 6: Module 5: Bivariate Analysis - Regression**

# Module 1: Basic Probability

# WHAT IS STATISTICS?


source: https://www.prosancons.com/education/pros-and-cons-of-statistics/

- science of changing your mind about default actions or prior beliefs with data
- scientific process of answering questions or making decisions with data
  - designing studies
  - collecting good data
  - describing the data with numbers and graphs
  - analyzing the data
  - making conclusions

credit: https://towardsdatascience.com/statistics-for-people-in-a-hurry-a9613c0ed0b

# TWO ASPECTS OF STATISTICS

1. Descriptive Statistics

2. Inferential Statistics

# DESCRIPTIVE STATISTICS

Objectives:

**Organize**

(sorting data, stemplots, frequency distributions of ungrouped data, ...)

**Summarize**

(frequency distributions of grouped data, measures of location, spread, skewness, …)

**Visualize**

(box plots, histograms, pie charts, …)

# DESCRIPTIVE STATISTICS: Summarizing Data – Single Number Summary

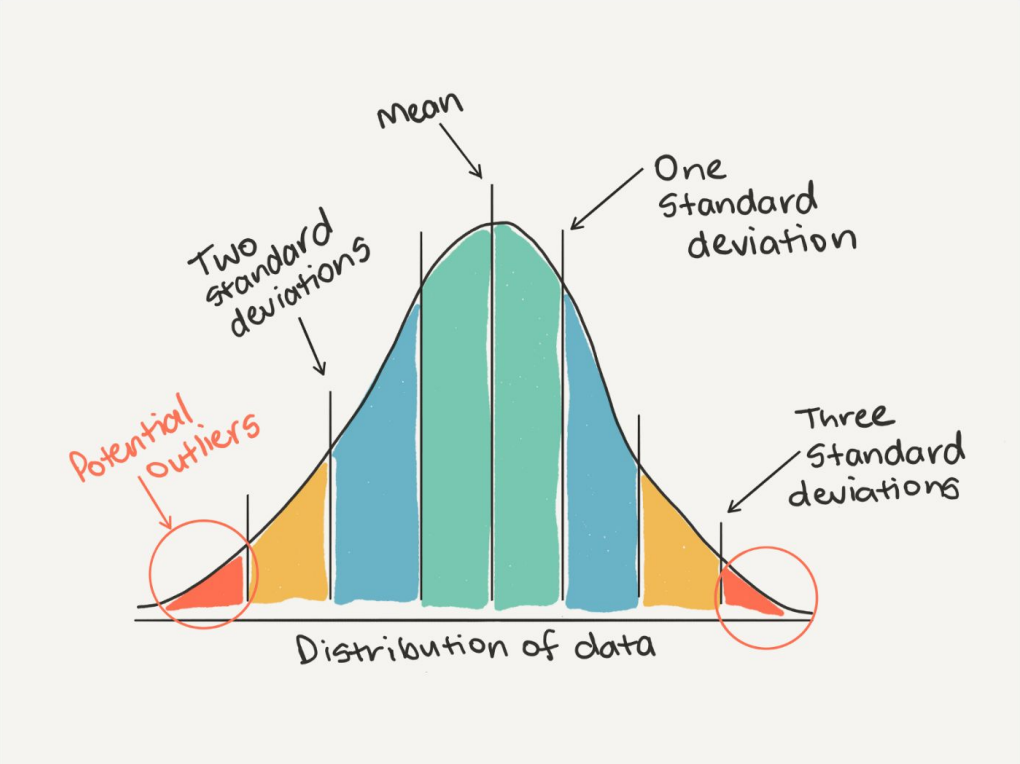
Objectives:

Obtain single number summary measures of

- Location (Central Tendency)
- Spread (Variation, Dispersion)

# Definition of terms: Central Tendency

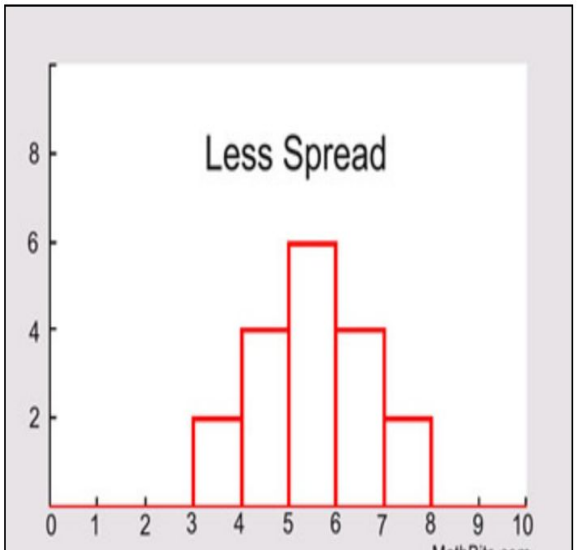
- This is a value that describes where most of the set of a data falls or clusters. Mean, Median, Mode are used to measure this.

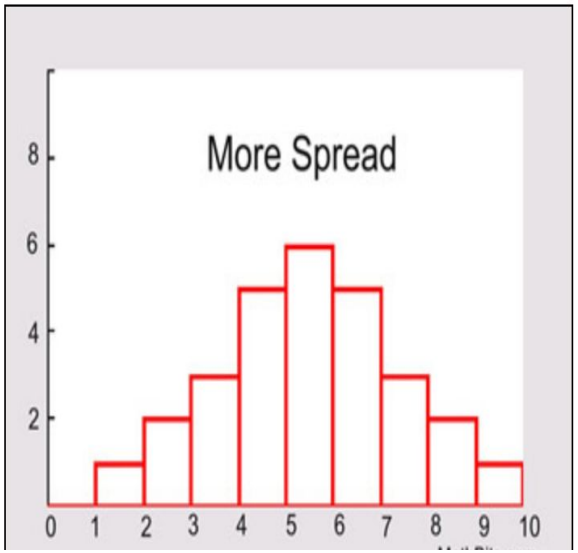



Source: Towards Data Science

# Definition of terms: Spread

It is an indicator of how far away from the center/mean/median data values are and is measured by variance, standard deviation, and interquartile range



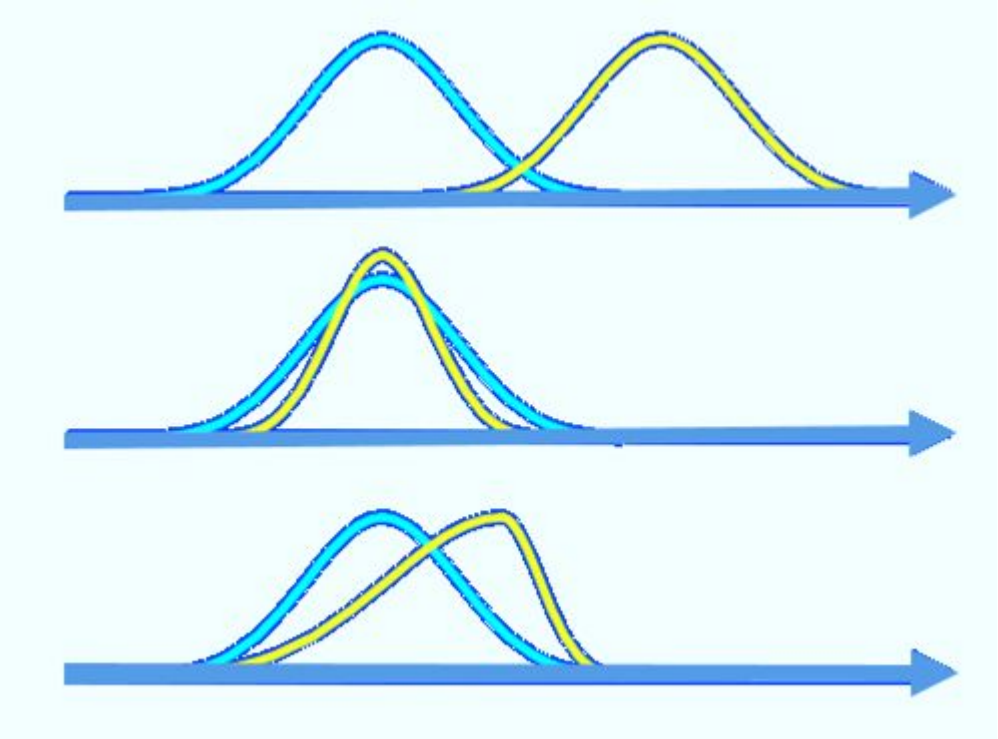

Source: https://mathbitsnotebook.com/

# Numerical Summaries of Data

Central Tendency
(LOCATION)

Variation
(DISPERSION)

Shape



Credit: http://tiny.cc/9ywe8y

# Central Tendency Measures

- Mean/Average
- Median
- Mode

# Mean/Average

- Sum of data points divided by Number of data points

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- A balance point for the distribution

Credit: http://bit.ly/2ZwNsn6

# Median

- Midpoint of the list if the measures are listed in ascending order
- The point on the measurement scale below which 50% of the score are located

Age of participants: 17 19 21 22 23 23 23 38

**Median = (22+23)/2 = 22.5**

**NOTE: IF THE NUMBER OF DATA POINTS IS EVEN, THE MEDIAN IS THE MEAN OF THE MIDDLE TWO NUMBERS**

Credit: http://bit.ly/2ZwNsn6

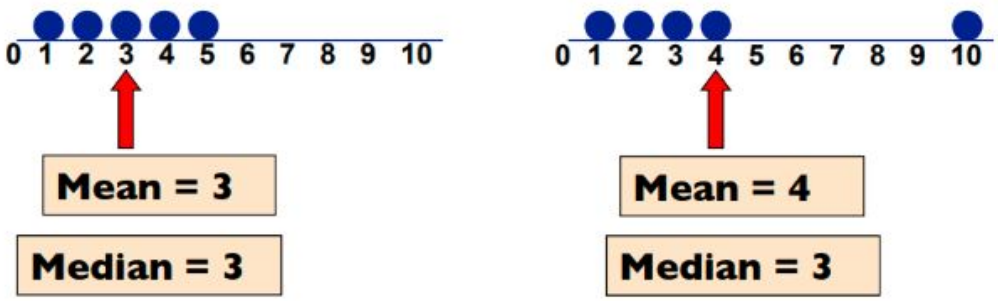
# Mode

- The value that has the greatest frequency
- It is not unique

Credit: http://bit.ly/2x2mbN6

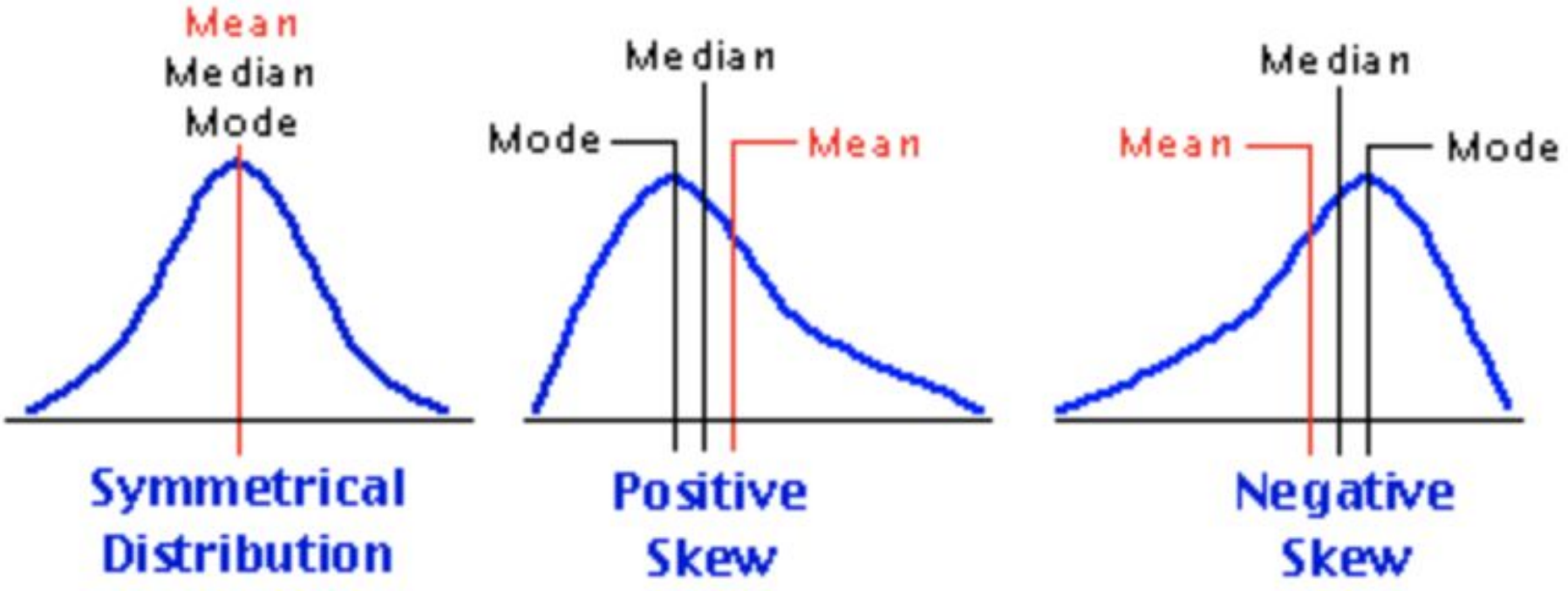
# Mean or Median ?

- MEAN is best for symmetric distributions without outliers
- MEDIAN is useful for skewed distributions or those with outliers




Credit: http://bit.ly/2ZwNsn6

# Skewed Data



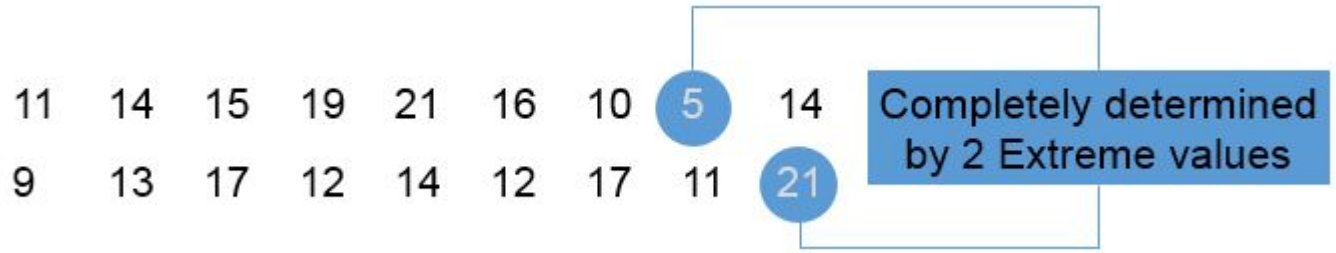

Credit: http://vassarstats.net/textbook/ch2pt1.html

# Variability Measures

- Range
- Percentiles
- Standard Deviation
- Variance

# Range

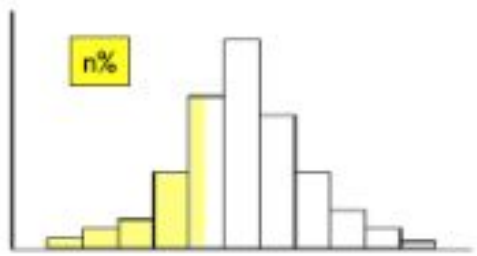
- Easy to compute and understand
- Non-Robust measure - influenced by extreme values
- Minimum value of the data set
- Maximum value of the data set




Credit: http://tiny.cc/u8iw7y

# Percentiles

- In general, the nth percentile is a value such that n% of the observations fall below it




$Q_1$ = 25th percentile

Median = 50th percentile

$Q_2$ = 75th percentile

Q1:= 25th percentile
**25% of data fall below this value**

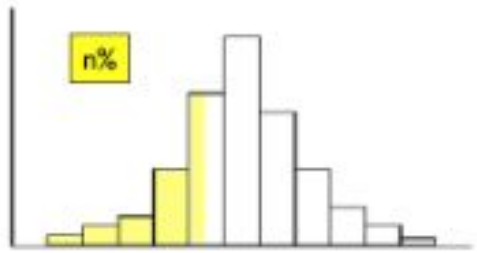
Q2 := 50th percentile
**50% of data fall below this value**

Q3 := 75th percentile
**75% of data fall below this value**

Credit: http://bit.ly/2ZwNsn6

# Percentiles

- In general, the nth percentile is a value such that n% of the observations fall below it




$Q_1$ = 25th percentile

Median = 50th percentile

$Q_2$ = 75th percentile

Q1:= 25th percentile
**Median to the left of the median**

Q2 := **MEDIAN**

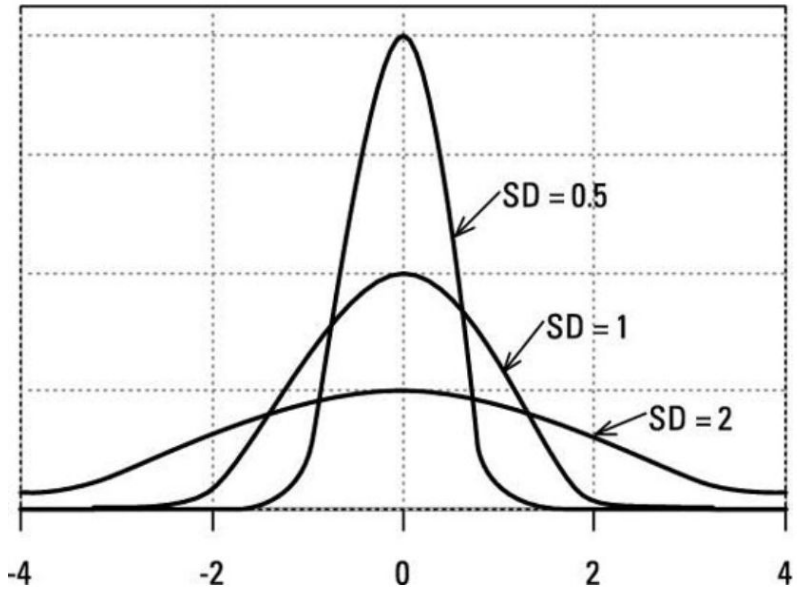
Q3 := 75th percentile
**Median to the right of the median**

Credit: http://bit.ly/2ZwNsn6

# Standard Deviation

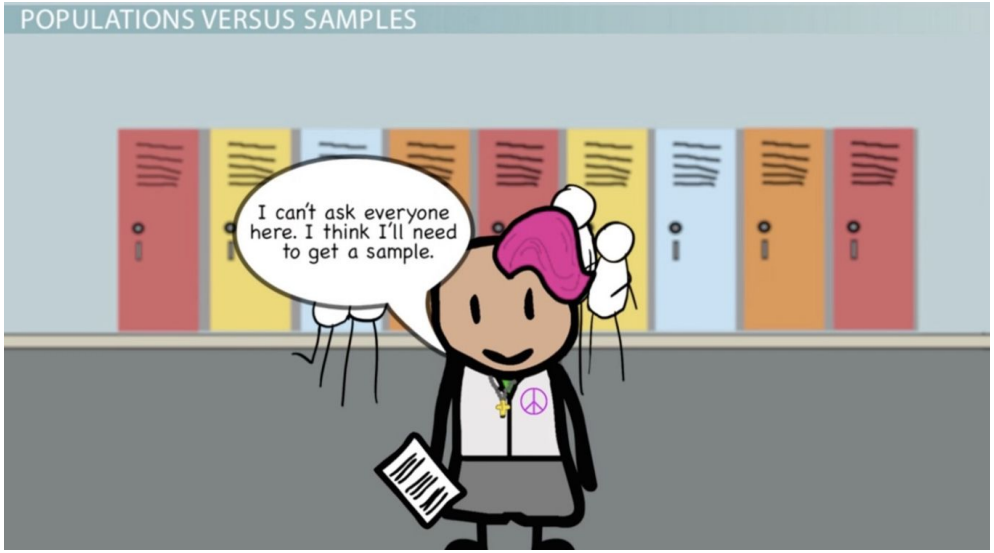
- Measures the spread from the mean
- Larger the variance/standard deviation larger the spread

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




Credit: http://bit.ly/2ZwNsn6

# Population and Sample



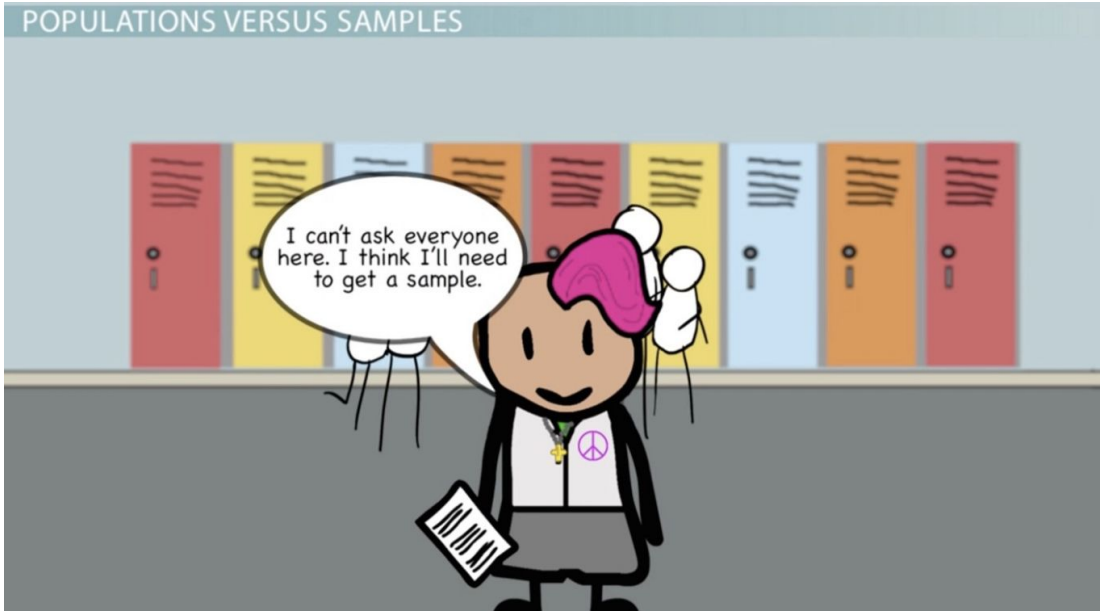

https://pharmstatistics.wordpress.com/author/pharmstatistics/

**Population:** entire group of individuals or objects about which we want information
Example:the population of Kathmandu

**Sample:** part of the population from which we actually collect information
Example:100 households picked randomly from Kathmandu

We use a **sample** to draw conclusions about the entire **population**.

# Population and Sample



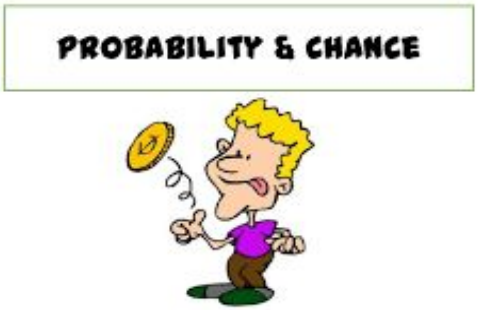

A **sampling design** describes exactly how to choose a sample from a population.

A **simple random sample (SRS)**is a sample where all the individuals as well as all the possible samples have an equal chance of getting selected.

# WHAT IS PROBABILITY?

- The weather forecasters say that there's a 60% chance of rain today.
- What are my chances of getting a job if I get a Master's degree?
- What are my chances of dying of a heart attack if I eat chocolate everyday?
- What are my chances of winning the lottery?
- What are the chances that the next person I meet on the street is a male?

Uncertainties!




Source: https://makeameme.org/meme/third-times-the-5afad1

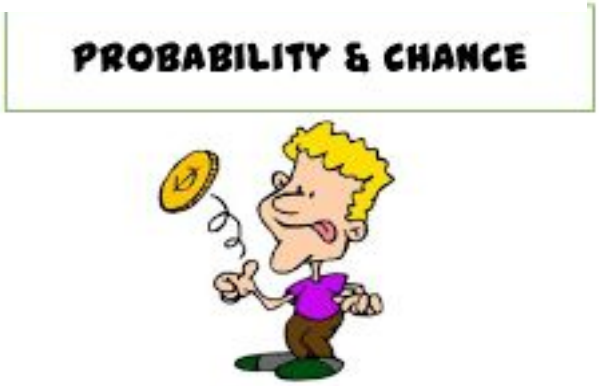
# WHAT IS PROBABILITY?

- Way to quantify uncertainties around events
- Uses numbers to tell us how likely something is to happens
- Probability or chance of something happening can be described by using words such as unlikely, impossible, likely, certainly etc.

Examples

- weather forecasting
- financial sector
- insurance companies
- development sector

Uncertainties!




Source: https://makeameme.org/meme/third-times-the-5afad1

# WHAT IS PROBABILITY?

## Theoretical Probability

Can be found without doing an experiment

## Experimental Probability

Can be found by repeating an experiment and observing the outcomes

# WHAT IS PROBABILITY?

## Theoretical Probability

Can be found without doing an experiment

## Experimental Probability

Can be found by repeating an experiment and observing the outcomes

Example: Coin flipping experiment - two possible outcomes “heads” or “tails”.

# WHAT IS PROBABILITY?

## Theoretical Probability

Can be found without doing an experiment

## Experimental Probability

Can be found by repeating an experiment and observing the outcomes

Example: Coin flipping experiment - two possible outcomes “heads” or “tails”.

1. The chance of getting a head is one in two.
2. so the probability of getting a head is ½.

# WHAT IS PROBABILITY?

## Theoretical Probability

Can be found without doing an experiment

## Experimental Probability

Can be found by repeating an experiment and observing the outcomes

Example: Coin flipping experiment - two possible outcomes “heads” or “tails”.

1. The chance of getting a head is one in two.
2. so the probability of getting a head is ½.

1. Toss a coin, record the outcome, and repeat the experiment many times (N)

# WHAT IS PROBABILITY?

## Theoretical Probability

Can be found without doing an experiment

## Experimental Probability

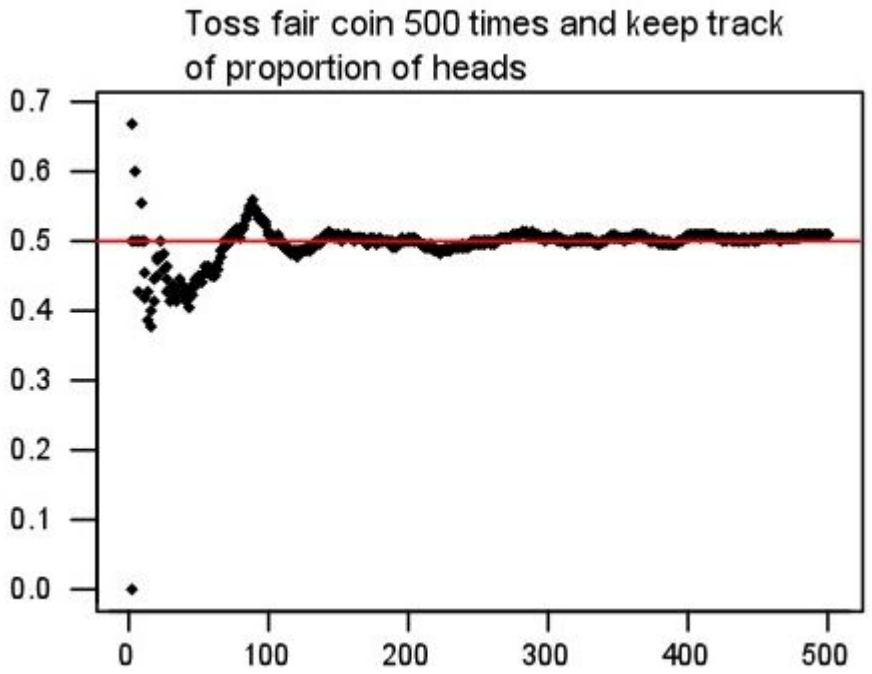
Can be found by repeating an experiment and observing the outcomes

Example: Coin flipping experiment - two possible outcomes “heads” or “tails”.

1. The chance of getting a head is one in two.
2. so the probability of getting a head is ½.

1. Toss a coin, record the outcome, and repeat the experiment many times (N).
2. Count the number of heads, n.

# WHAT IS PROBABILITY?

# WHAT IS PROBABILITY?

## Theoretical Probability

Can be found without doing an experiment

## Experimental Probability

Can be found by repeating an experiment and observing the outcomes

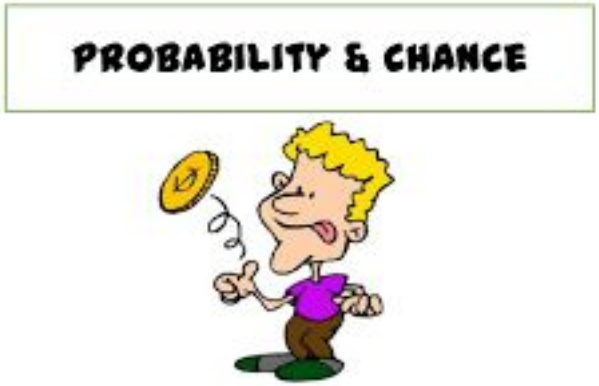
Example: Coin flipping experiment - two possible outcomes “heads” or “tails”.

1. The chance of getting a head is one in two.
2. so the probability of getting a head is ½.

1. Toss a coin, record the outcome, and repeat the experiment many times (N)
2. Count the number of heads, n
3. Probability of getting a head = n/N

# WHAT IS PROBABILITY?

- Let's flip a coin - two possible outcomes "heads" or "tails"
- Outcome is not known in advance
- Repeat the coin tossing experiment many times - we will see a pattern
- A remarkable fact about chance behavior - it is unpredictable in the short run but has a regular and predictable pattern in the long run. This is the basis for the idea of probability.



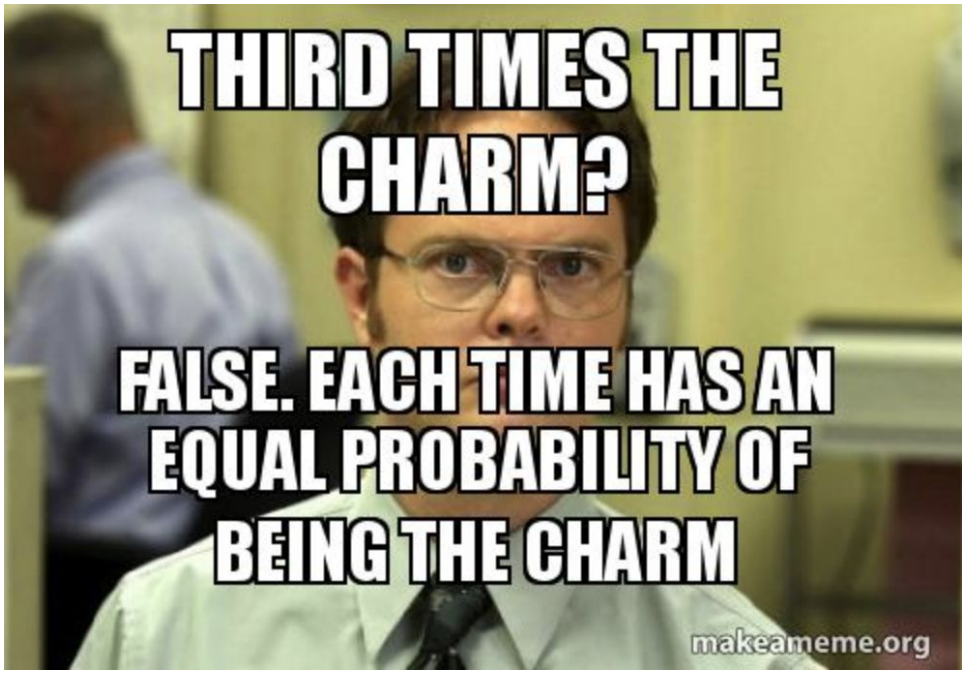

Source: https://makeameme.org/meme/third-times-the-5afad1

# WHAT IS PROBABILITY?

Probability is a mathematical framework that helps us measure the "likelihood" of results or outcome of stochastic or **random events**.

Random Events

- events whose outcomes are not known in advance
- however, a regular distribution of outcomes in a large number of repetitions




Source: https://makeameme.org/meme/third-times-the-5afad1

# WHAT IS PROBABILITY?

In order to define probability, we need to conduct an *experiment* and figure out the *event* or outcome of interest.

If A is the event we are interested in, then the probability of A is denoted by P(A). Formally,

$$\text{Probability of an outcome P(A)} = \frac{\text{Number of favorable outcomes (in the event)}}{\text{Number of all possible outcomes}}$$

# WHAT IS PROBABILITY?

1. **Define the *experiment*:** What are you doing to create outcomes; e.g., flipping a coin
2. **List all possible *outcomes*:** With a coin, we'll get either heads or tails. This is called the *sample space S*

$$S = \{\text{heads, tails}\}$$

3. **Determine probabilities of all outcomes from Step 2:** If the coin is fair, the probability of a head will be 50% as will the probability of a tail
4. **Determine event(s) we are interested in:** Let's observe the number (or percentage) of heads that we get for instance
5. The sum of the probabilities of the event(s) you are interested in = probability that the event(s) will occur

# WHAT IS PROBABILITY?

Example: Let's roll a die.

# WHAT IS PROBABILITY?

Example: Let's roll a die.

A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6.

# WHAT IS PROBABILITY?

Example: Let's roll a die.

A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6,

and if the die is fair, the probability of each event occurring is $\frac{1}{6}$

# WHAT IS PROBABILITY?

Example: Let's roll a die.

A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6,
and if the die is fair, the probability of each event occurring is ⅙

What is the probability of rolling an odd number?

# WHAT IS PROBABILITY?

Example: Let's roll a die.

A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6,
and if the die is fair, the probability of each event occurring is ⅙

What is the probability of rolling an odd number?

A: There are 3 odd numbers between 1 and 6, they are 1,3,5

Number of ways we can get an odd number is 3.

Number of all possible outcomes is 6.

P(odd number) = 3/6= 0.5

# WHAT IS PROBABILITY?

Example: Let's take an example of rolling a die.

A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6,

and if the die is fair, the probability of each event occurring is ⅙.

Q1: What is the probability of rolling an odd number?

A: There are 3 odd numbers between 1 and 6, they are 1,3,5, hence,

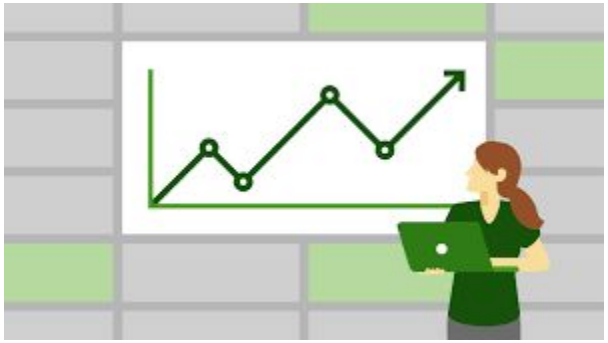
Number of ways we can get an odd number is 3.

Number of all possible outcomes is 6.

Hence,

P(odd number) = 3/6= 0.5

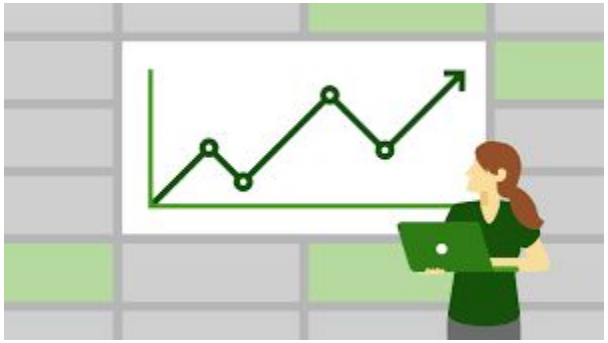
# Properties/Rules of Probability



1. The probability of any event A is always between 0 and 1.

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

# Properties/Rules of Probability



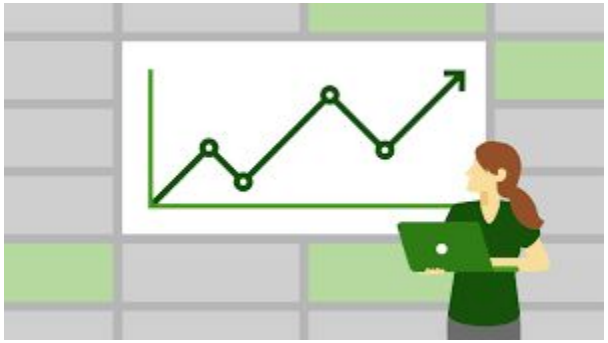
1. The probability of any event A is always between 0 and 1.

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

A probability of 0 means impossibility; i.e., the event will never happen.

A probability of 1 means certainty; i.e., the event will always happen.

# Properties/Rules of Probability



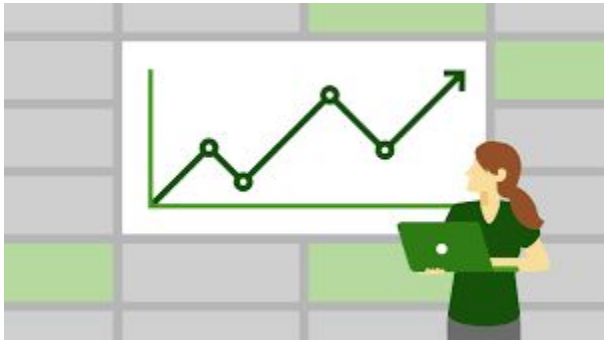
1. The probability of any event A is always between 0 and 1.

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

2. The sum of probabilities of all the possible outcomes (A, B, C …) should be equal to 1.

$$P(A) + P(B) + P(C) + \ldots = 1$$

# Properties/Rules of Probability



1. The probability of any event A is always between 0 and 1.

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

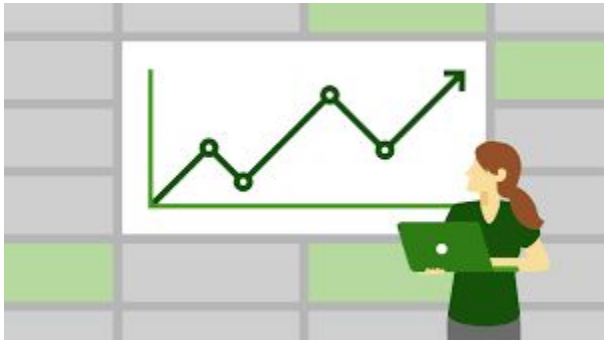
2. The sum of probabilities of all the possible outcomes (A, B, C …) should be equal to 1.

$$P(A) + P(B) + P(C) + \ldots = 1$$

3. The probability that an event A does not occur (the complement probability)

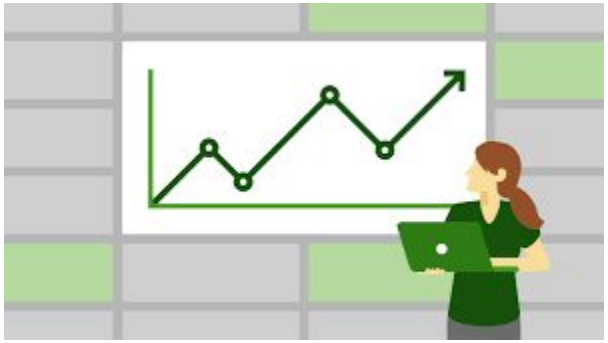
$$P(A \text{ does not occur}) = 1 - P(A)$$

# EXERCISE



Let’s toss two coins. What is the sample space?

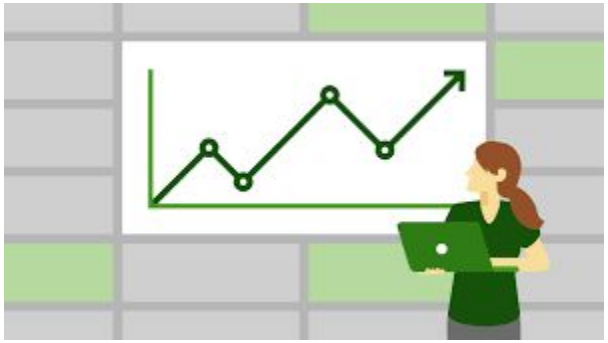
# EXERCISE



Let’s toss two coins. What is the sample space?

S = {HH, HT, TH, TT}, H~heads, T~tails

# EXERCISE

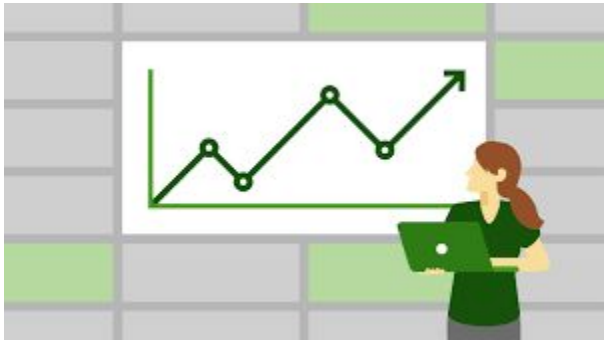


Let’s toss two coins. What is the sample space?

S = {HH, HT, TH, TT}, H~heads, T~tails

What is the probability of getting 2 heads?

# EXERCISE



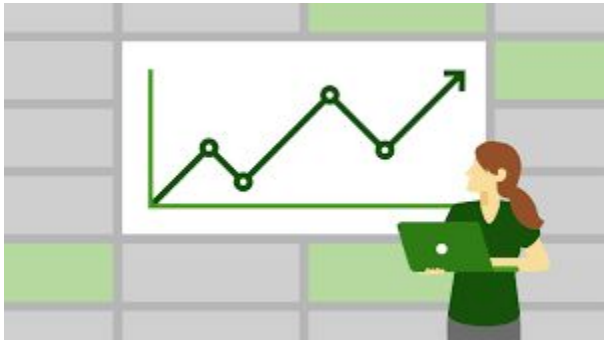
Let’s toss two coins. What is the sample space?

S = {HH, HT, TH, TT}, H~heads, T~tails

What is the probability of getting 2 heads?

P(HH) = 1/4

# EXERCISE

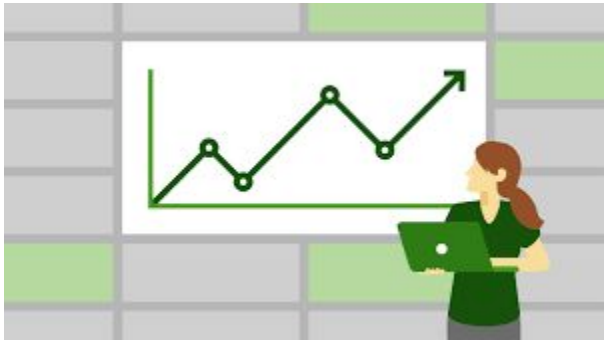


Let's toss two coins. What is the sample space?

S = {HH, HT, TH, TT}, H~heads, T~tails

What is the probability of getting 1 head and 1 tail? (order does not matter)

# EXERCISE



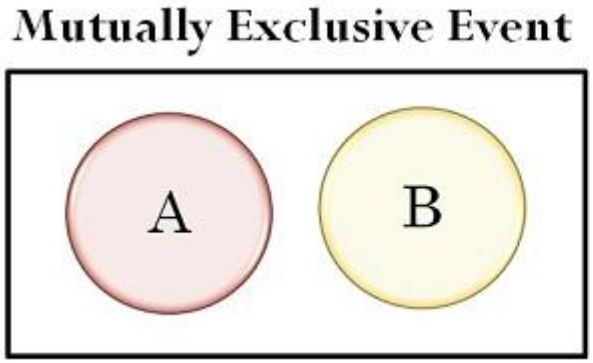
Let's toss two coins. What is the sample space?

S = {HH, HT, TH, TT}, H~heads, T~tails

What is the probability of getting 1 head and 1 tail? (order does not matter)

P(1H and 1T) = 2/4

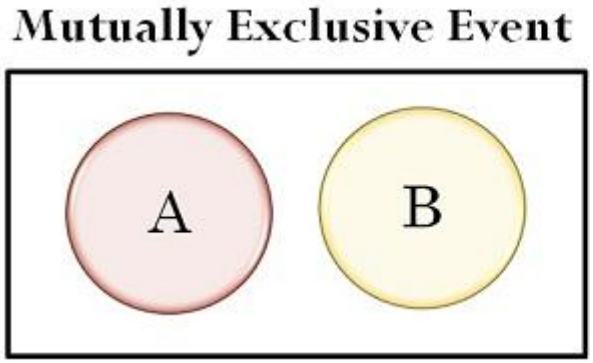
# Mutually Exclusive Events




source: https://keydifferences.com/difference-between-mutually-exclusive-and-independent-events.html

When events A and B cannot occur at the same time, they are called **mutually exclusive events.**

# Mutually Exclusive Events

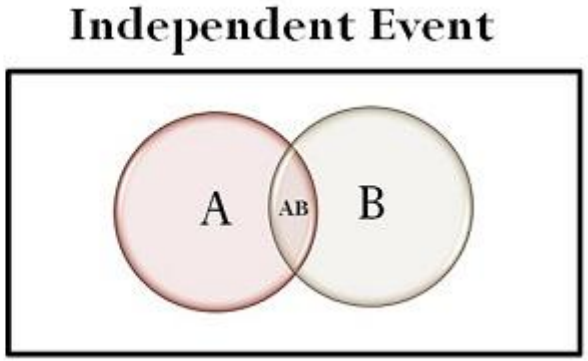



source: https://keydifferences.com/difference-between-mutually-exclusive-and-independent-events.html

When events A and B cannot occur at the same time, they are called **mutually exclusive events.**

- A person cannot be from Province 1 and Province 2 at the same time
- When we toss a coin, we cannot get a head and a tail at the same time

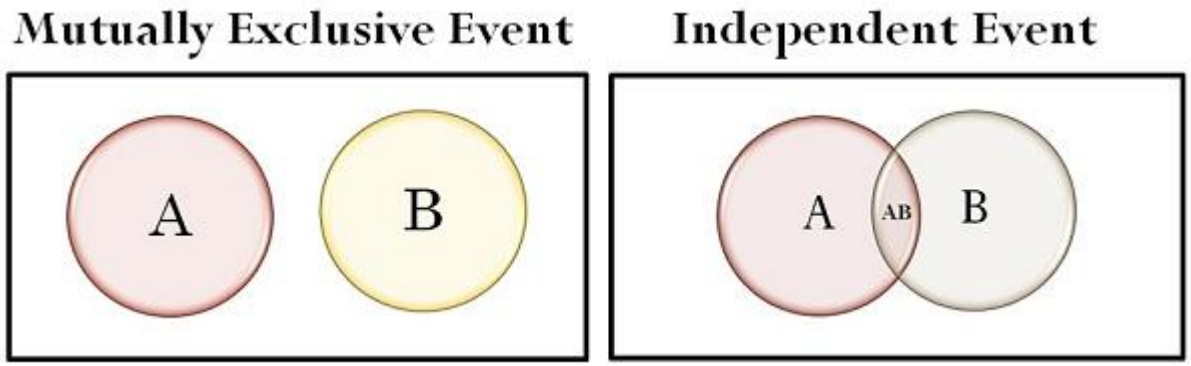
# Independent Events




source: https://keydifferences.com/difference-between-mutually-exclusive-and-independent-events.html

Event A and B are called **independent events** if event B is unaffected by event A, and vice versa.

- When we toss 2 coins, what we get in toss 2 is not affected by what we get in the toss 1.
- Fertility rate in Nepal does not affect the probability of an earthquake in Nepal




source: https://keydifferences.com/difference-between-mutually-exclusive-and-independent-events.html

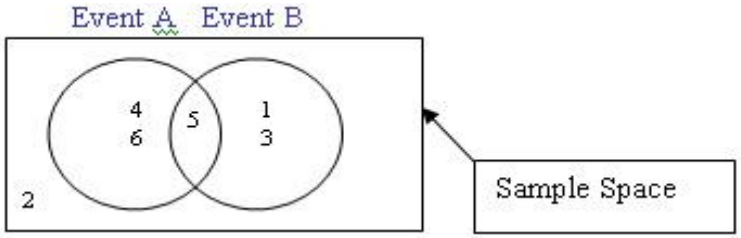
## Mutually exclusive events and independent events

When events A and B cannot occur at the same time, they are called **mutually exclusive events.**
If event B is unaffected by event A, they are called **independent events.**

# Compound Probabilities

Example: A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6, and if the die is fair, the probability of each event occurring is ⅙. We might be interested in the following events -

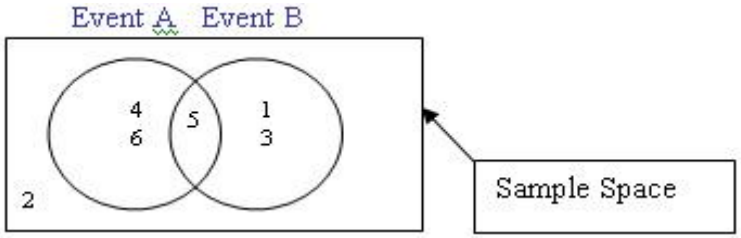
- Event A: Rolling a 4 or higher (4,5,6). denoted as P(4 or 5 or 6)
- Event B: Rolling an odd number (1,3,5).



# Compound Probabilities

Example: A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6, and if the die is fair, the probability of each event occurring is ⅙. We might be interested in the following events -

- Event A: Rolling a 4 or higher (4,5,6). denoted as P(4 or 5 or 6)
- Event B: Rolling an odd number (1,3,5).




What is the probability of rolling a 4 or higher OR rolling an odd number?

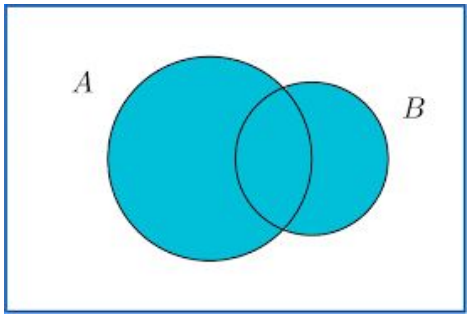
What is the probability rolling a 4 or higher AND rolling an odd number?

# Compound Probabilities

Two kinds of compound events -

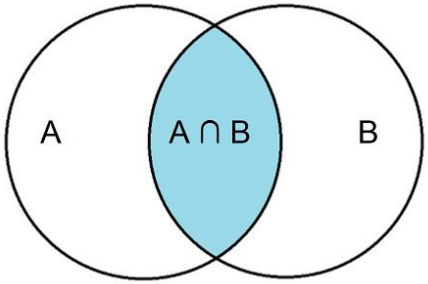
1. The **union event of A and B** is the event of A or B or both happening.

   **P(A or B) or P(A∪B)**




2. The **intersection event of A and B** is the event of both A and B happening.

   **P(A and B) or P(A∩B)**

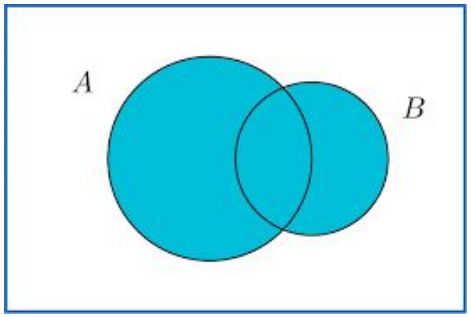


# Compound Probabilities

Two kinds of compound events -

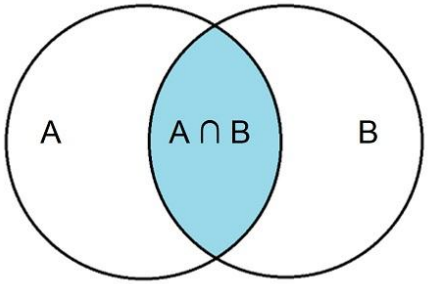
1. The **union event of A and B** is the event of A or B or both happening.

   **P(A or B) or P(A∪B)**




2. The **intersection event of A and B** is the event of both A and B happening.

   **P(A and B) or P(A∩B)**




**Addition Rule for any two events A and B**

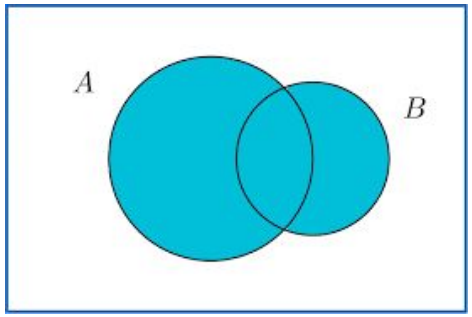
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

# Compound Probabilities

Two kinds of compound events -

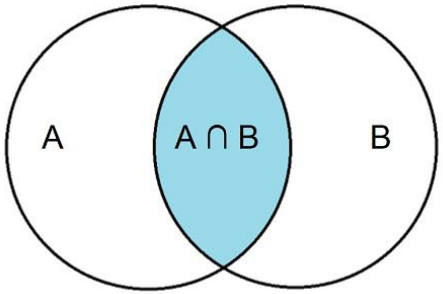
1. The **union event of A and B** is the event of A or B or both happening.

   **P(A or B) or P(A∪B)**




2. The **intersection event of A and B** is the event of both A and B happening.

   **P(A and B) or P(A∩B)**




**Addition Rule for any two events A and B**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

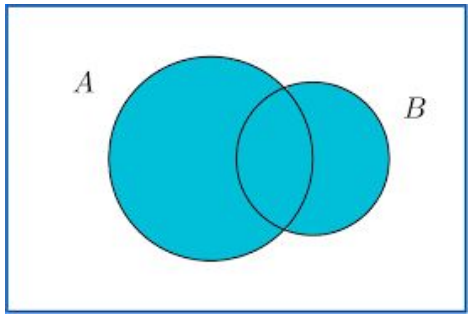
**Mutually Exclusive/ Disjoint Events:** $P(A \cap B) = 0$

# Compound Probabilities

Two kinds of compound events -

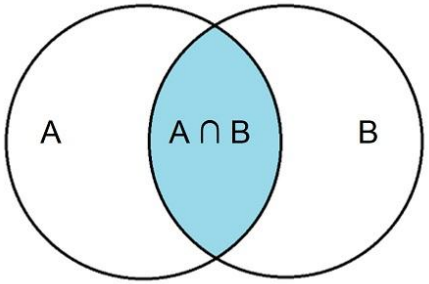
1. The **union event of A and B** is the event of A or B or both happening.

   **P(A or B) or P(A∪B)**




2. The **intersection event of A and B** is the event of both A and B happening.

   **P(A and B) or P(A∩B)**




**Addition Rule for any two events A and B**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Mutually Exclusive/ Disjoint Events:** $P(A \cap B) = 0$

**Independent Events:** $P(A \cap B) = P(A)P(B)$

# Compound Probabilities

**Addition Rule for any two events A and B**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Addition Rule for Mutually Exclusive events**

$$P(A \cup B) = P(A) + P(B)$$

**Addition Rule for Independent events**

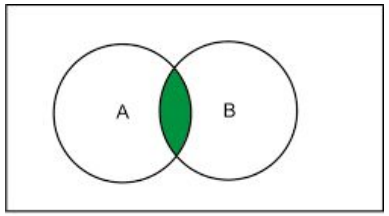
$$P(A \cup B) = P(A) + P(B) - P(A)P(B)$$

# Compound Probabilities

Two kinds of compound events -

1. The union event of A and B is the event of A or B or both happening. The probability of union of A and B is denoted by P(A or B) or P(A∪B)

2. The intersection event of A and B is the event of both A and B happening.The probability of intersection of A and B is denoted by P(A and B) or P(A∩B).




A AND B

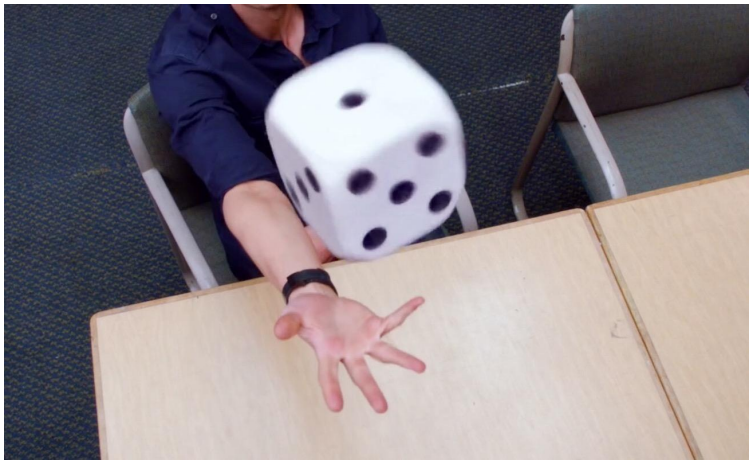
A∩B

**Multiplication Rule for independent events A and B**

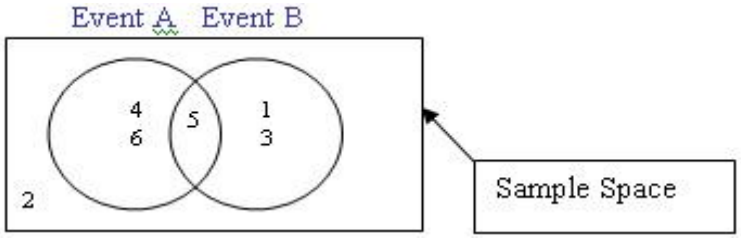
$$P(A \cap B) = P(A)P(B)$$

# Compound Probabilities

Example: A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6, and if the die is fair, the probability of each event occurring is ⅙. We might be interested in the following events -



- Event A: Rolling a 4 or higher (4,5,6). denoted as P(4 or 5 or 6)
- Event B: Rolling an odd number (1,3,5).

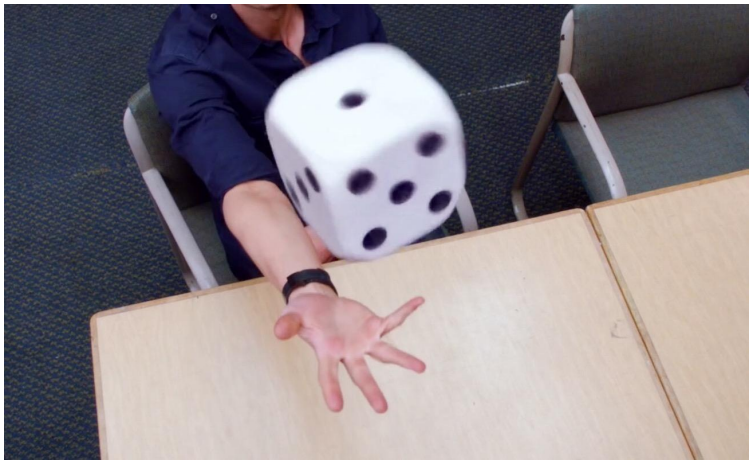



P(rolling a 4 or higher OR rolling an odd number) = P(A∪B)

P(rolling a 4 or higher AND rolling an odd number) = P(A∩B)

# Compound Probabilities

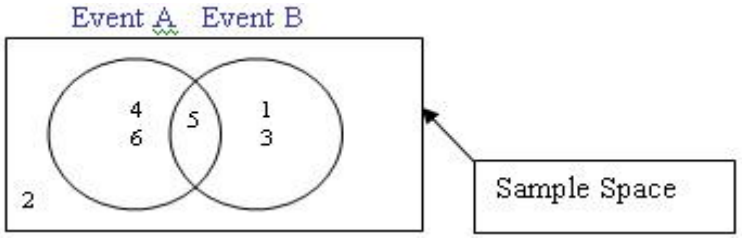
Example: A die with six sides has six possible outcomes - 1, 2, 3, 4, 5 or 6, and if the die is fair, the probability of each event occurring is ⅙. We might be interested in the following events -



- Event A: Rolling a 4 or higher (4,5,6). denoted as P(4 or 5 or 6)
- Event B: Rolling an odd number (1,3,5).

P(rolling a 4 or higher OR rolling an odd number) = P(A∪B)

P(rolling a 4 or higher AND rolling an odd number) = P(A∩B)




$$P(A) = \frac{3}{6}, \; P(B) = \frac{3}{6},$$

$$P(A \cap B) = \frac{1}{6}$$

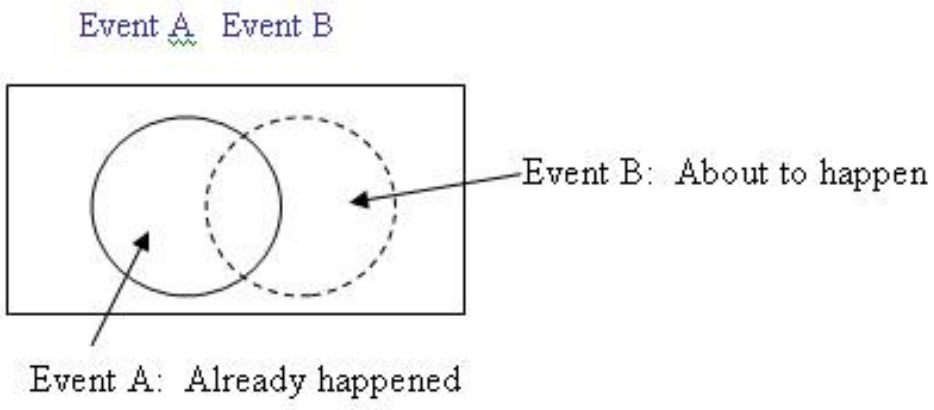
Hence,

$$P(A \cup B) = \frac{3}{6} + \frac{3}{6} - \frac{1}{6} = \frac{5}{6}$$

# Conditional Probability

Suppose the Government of Nepal is interested in understanding if higher pay is related to graduate degree. They want to know the probability of getting a higher salary if (or given that) you have a graduate degree.

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

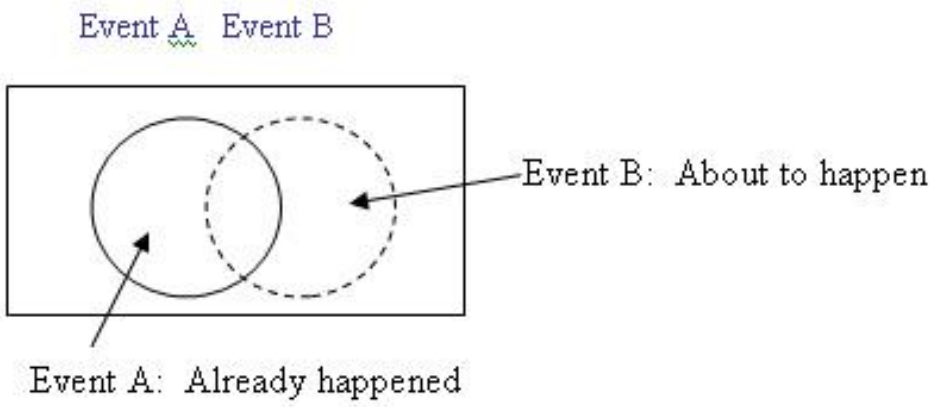


# Conditional Probability

Suppose the Government of Nepal is interested in understanding if higher pay is related to graduate degree. They want to know the probability of getting a higher salary if (or given that) you have a graduate degree.

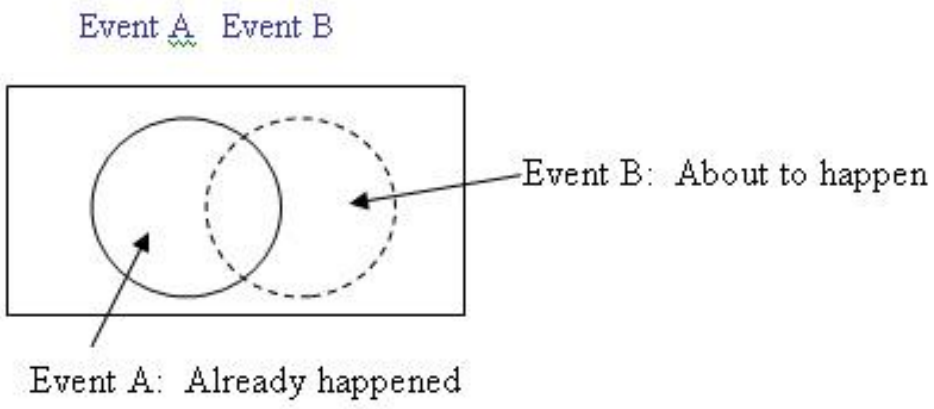
**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B given Event A) = ?



# Conditional Probability

Assume Event A is that you have your graduate degree, and that Event B is that you get higher pay upon graduation. We can express the conditional probability, P(B|A), as the probability of getting a higher salary if (or given that) you have a graduate degree. Pictorially, we can draw this as shown below. Note that Event A, having your graduate degree, has already happened. Event B, getting a higher salary, is about to happen.



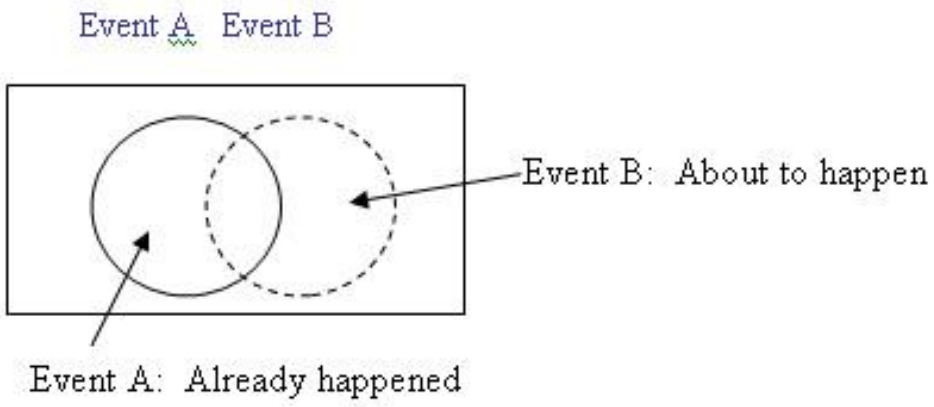
# Conditional Probability

**A conditional probability denotes the probability of an event B occurring given that event A has occurred, and is denoted by P(B|A).**

already happened

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(Event B | Event A) = ?



# Conditional Probability

**A conditional probability denotes the probability of an event B occurring given that event A has occurred, and is denoted by P(B|A).**

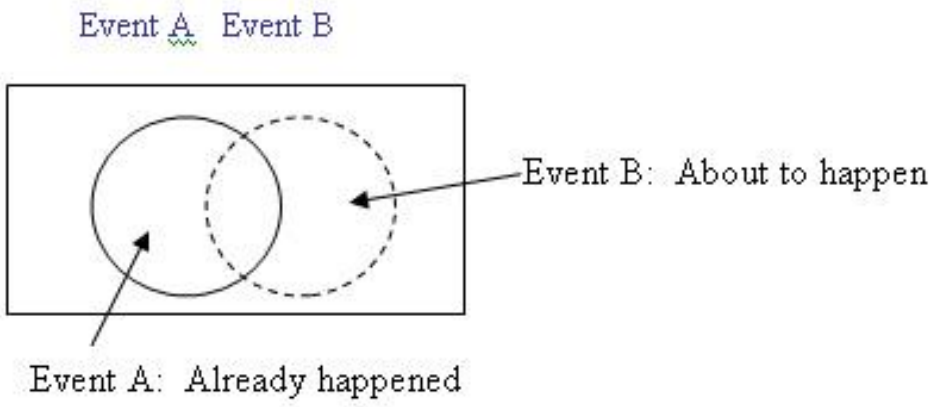
already happened

**Event A:** you have your graduate degree

**Event B:** you get higher pay upon graduation

about to happen

P(Event B | Event A) = ?



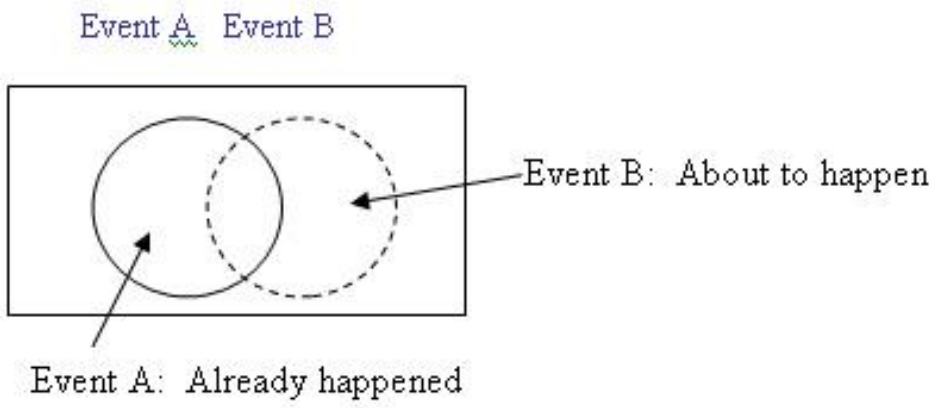
# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

**NOTE:**

P(Event B | Event A) ≠ P(Event B and Event A)



# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

Rather, we want to find the number of those with higher degrees and higher salaries—the intersection—but only as a fraction of those with higher degrees.

Conditional probabilities have a reduced sample space—we only care about results based on what has already happened, and nothing else!

**P(A) := Probability of getting a graduate degree**
**P(B) := Probability of getting a higher pay**

# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

Rather, we want to find the ratio of those with higher degrees and higher salaries—the intersection—but only as a fraction of those with higher degrees.

Conditional probabilities have a reduced sample space—we only care about results based on what has already happened, and nothing else!

**P(A) := Probability of getting a graduate degree**
**P(B) := Probability of getting a higher pay**

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \; P(A) > 0$$

# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

**P(A) := Probability of getting a graduate degree**
**P(B) := Probability of getting a higher pay**

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

# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

**P(A) := Probability of getting a graduate degree**
**P(B) := Probability of getting a higher pay**

**NOTE:**

**P(B|A) ≠ P(A|B)**

**P(higher pay | grad degree) ≠**
**P(grad degree | higher pay)**

# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

**What if: the probability of getting a higher salary doesn't seem to depend in any way on having a graduate degree?**

# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

**What if: the probability of getting a higher salary doesn't seem to depend in any way on having a graduate degree?**

**P(B|A) = ?**

# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

**What if: the probability of getting a higher salary doesn't seem to depend in any way on having a graduate degree?**

**P(B|A) = P(B)**

**P(higher pay | grad degree) = P(higher pay)**

# Conditional Probability

**Event A:** you have your graduate degree
**Event B:** you get higher pay upon graduation

P(higher pay given you have a graduate degree) =
P(Event B | Event A) = ?

**What if: the probability of getting a higher salary doesn't seem to depend in any way on having a graduate degree?**

**P(B|A) = P(B)**

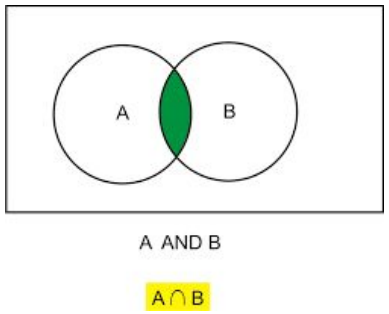
**P(higher pay | grad degree) = P(higher pay)**

**getting a higher pay and graduate degree are independent events**

# Recall Independent Events

If event B is unaffected by event A, they are called **independent events**. That means

$$P(B|A) = P(B)$$




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

**Multiplication Rule for independent events A and B**

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

# Example

Assume we work for a company and we are tracking shopping behavior for one of our products. We have men and women as customers, and our product is packaged in two different colors. We would like to see if there is any association between the shopper's gender and the color of the product packaging.
The table below shows purchase results from a recent in-store survey:

| | Men | Women | Totals |
|---|---|---|---|
| Blue Packaging | 20 | 5 | 25 |
| Pink Packaging | 10 | 30 | 40 |
| Totals | 30 | 35 | 65 |

# Example

Assume we work for a company and we are tracking shopping behavior for one of our products. We have men and women as customers, and our product is packaged in two different colors. We would like to see if there is any association between the shopper's gender and the color of the product packaging.
The table below shows purchase results from a recent in-store survey:

| | Men | Women | Totals |
|---|---|---|---|
| Blue Packaging | 20 | 5 | 25 |
| Pink Packaging | 10 | 30 | 40 |
| Totals | 30 | 35 | 65 |

**Are the shopper's gender and color of the product packaging dependent?**

# Example

| | Men | Women | Totals |
|---|---|---|---|
| Blue Packaging | 20 | 5 | 25 |
| Pink Packaging | 10 | 30 | 40 |
| Totals | 30 | 35 | 65 |

**Are the shopper's gender and color of the product packaging dependent?**

**?**

$$P(\text{Pink}|\text{Male}) = P(\text{Pink})$$

# Example

| | Men | Women | Totals |
|---|---|---|---|
| Blue Packaging | 20 | 5 | 25 |
| Pink Packaging | 10 | 30 | 40 |
| Totals | 30 | 35 | 65 |

**Are the shopper's gender and color of the product packaging dependent?**

**?**

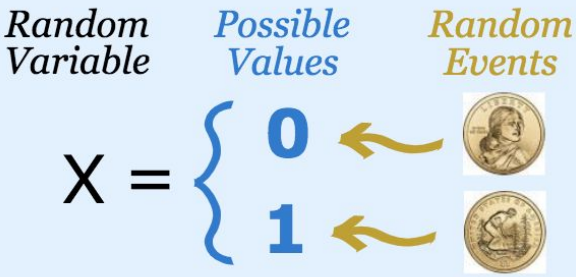
$$P(\text{Pink}|\text{Male}) = P(\text{Pink})$$

# Random variable

A variable X whose value is a numerical outcome of a statistical experiment

Example: Tossing a coin: we could get Heads or Tails.

Let's give them the values **Heads=0** and **Tails=1** and we have a Random Variable "X":

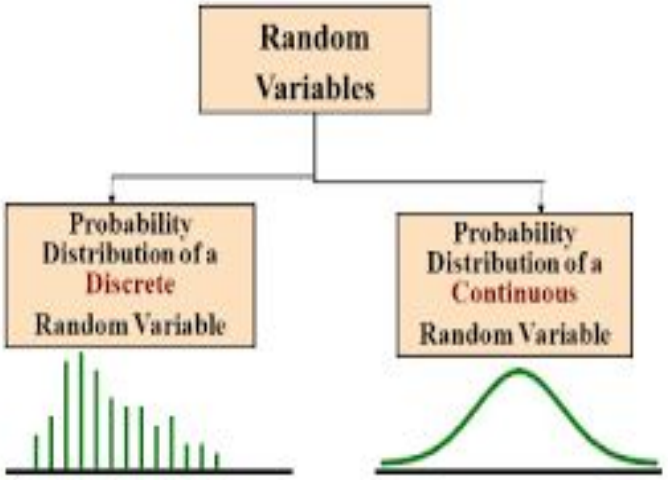



In short:

$$X = \{0, 1\}$$

source: https://www.mathsisfun.com/data/random-variables.html

# Random variable




## Random Variables

- Discrete Random Variables:

  a random variable that can assume only a countable number of values. The value of a discrete random variable comes from counting.

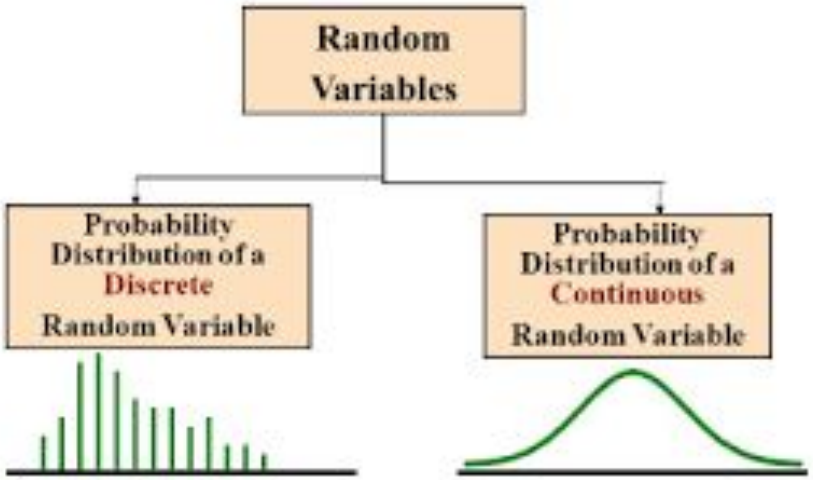
- Continuous Random Variable:

  random variables that can assume any value on a continuum. Measurement is required to determine the value for a continuous random variable.

# Expected Value of a Random Variable

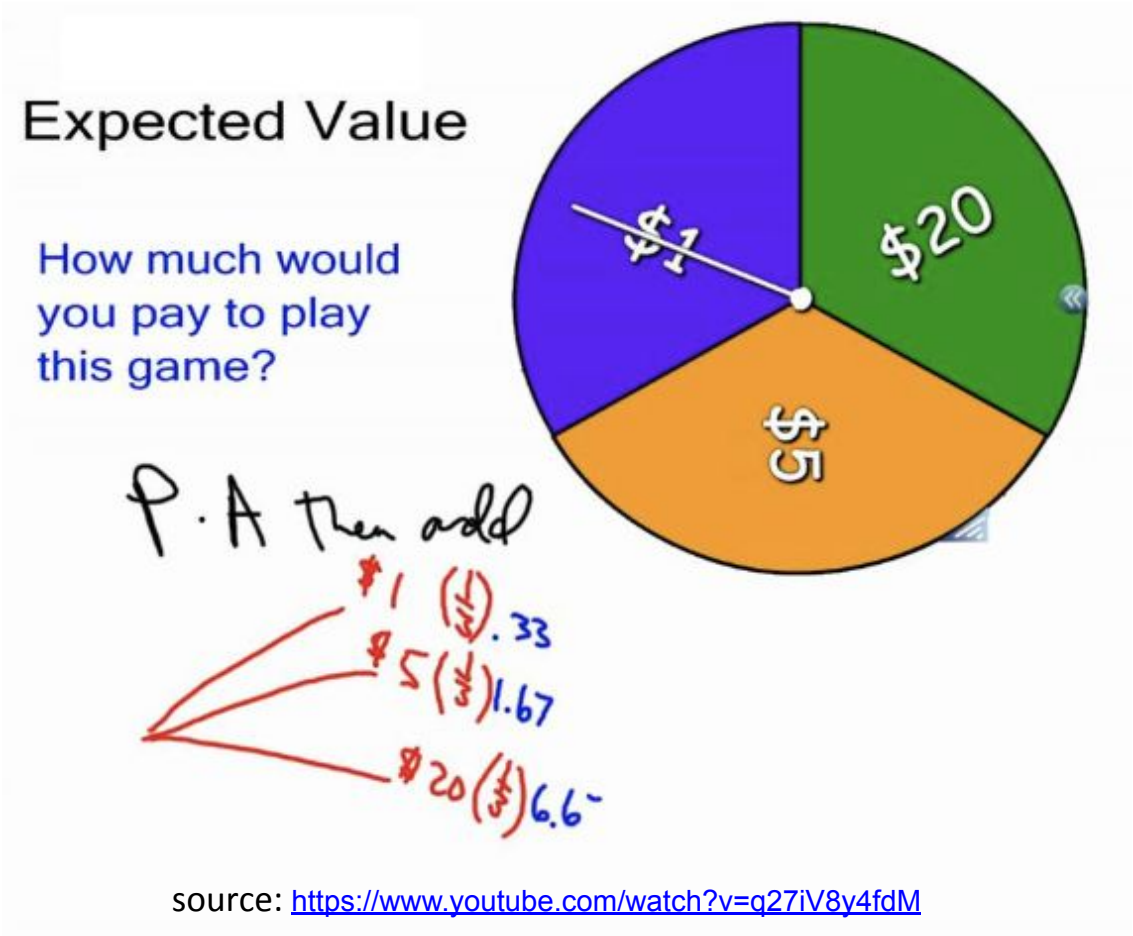
Definition: If X is a random variable ( $x_1, x_2, x_3 \ldots x_n$ ) that describes the outcomes of an experiment E, and $P(x_1), P(x_2), P(x_3) \ldots P(x_n)$ are their probabilities, then the **Expected Value (E(X))** of the random variable X is given by

$$E(X) = x_1 * P(x_1) + x_2 * P(x_2) + x_3 * P(x_3) \ldots x_n * P(x_n)$$



# Expected Value of a Random Variable

Example



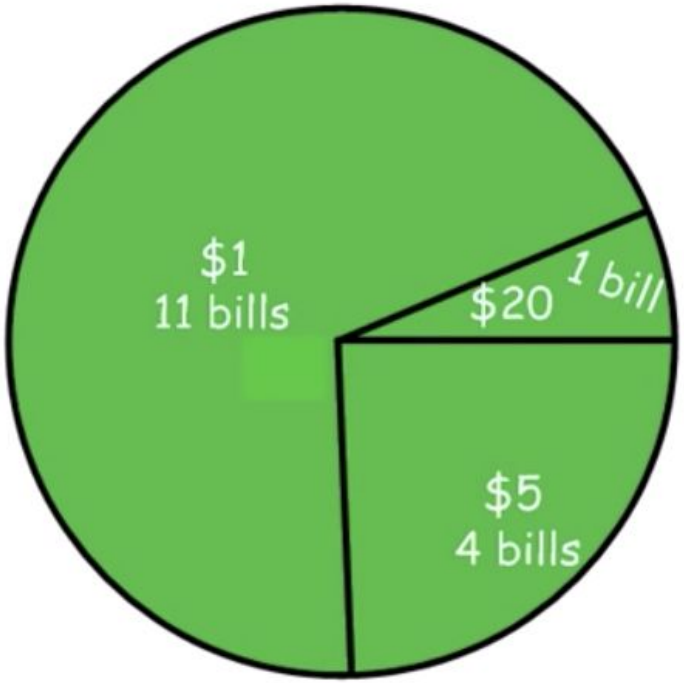

source: https://www.youtube.com/watch?v=q27iV8y4fdM

# Expected Value of a Random Variable

Example

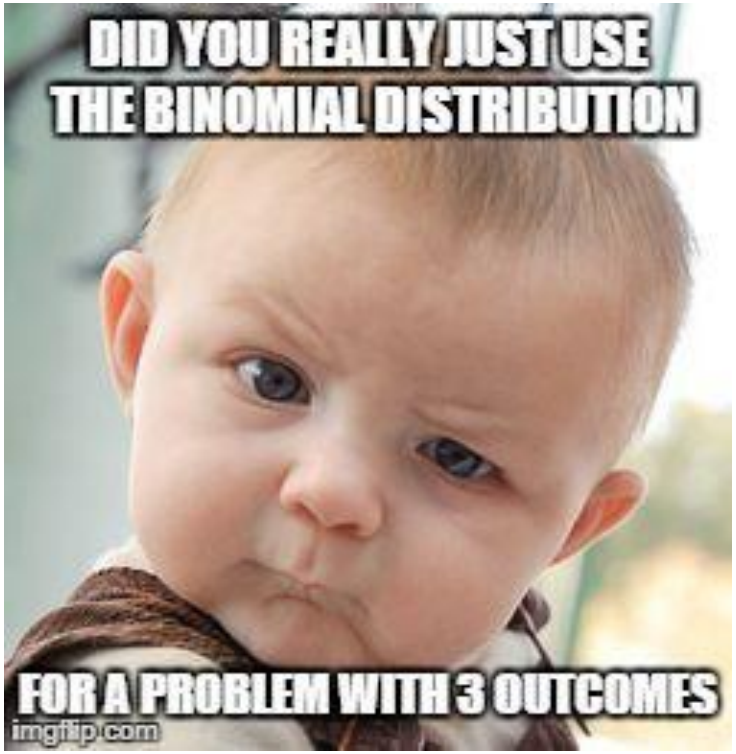
12.4 Random Variables and Expected Value

Suppose your wallet contains eleven $1 bills, four $5 bills, and one $20 bill. Immagine that you reach into your wallet and remove a bill at random, and then replace it. If you were to do this many times, what is the average value of the bill you remove?




source: https://www.youtube.com/watch?v=q27iV8y4fdM

# Module 2: Binomial, Normal Distribution, Statistical Inference




source: https://imgflip.com/

# Binomial Distribution

- A type of distribution that has two possible outcomes (the prefix “bi” means two, or twice) E.g., success or failure, hitting a red light or not, developing a side effect or not
- A random variable (x) has a binomial distribution if all of the following conditions are met:
  1. There are a fixed number of trials (n).
  2. Each trial has two possible outcomes: success or failure.
  3. The probability of success (p) is the same for each trial.
  4. The trials are independent, meaning the outcome of one trial doesn’t influence that of any other.

# Binomial Distribution

Example: You flip a fair coin 100 times and count the number of heads. Does this represent a binomial random variable?

1. **Are there a fixed number of trials?**
   Yes, you're flipping the coin 100 times, n = 100.
2. **Does each trial have only two possible outcomes, success or failure?**
   The outcome of each flip is either heads or tails, flipping a head represents success and flipping a tail is a failure.
3. **Is the probability of success the same for each trial?**
   Because the coin is fair, the probability of success is $p = 1/2$ for each trial.
4. **Are the trials independent?**
   We assume the coin is being flipped the same way each time, which means the outcome of one flip doesn't affect the outcome of subsequent flips.

# Binomial Distribution

The **Expected Value** of a binomial variable is

$$\mu = np$$

where,

n = number of events

p = probability

and the **Standard Deviation** is

$$\sigma = \sqrt{np(1-p)}$$

# Binomial Distribution

The Expected Value of a binomial variable is

$$\mu = np$$

Example of coin flip

n = 100

p = 0.5

$\mu = 100 * 0.5 = 50$

and the Standard Deviation is

$$\sigma = \sqrt{np(1-p)}$$

Example of coin flip

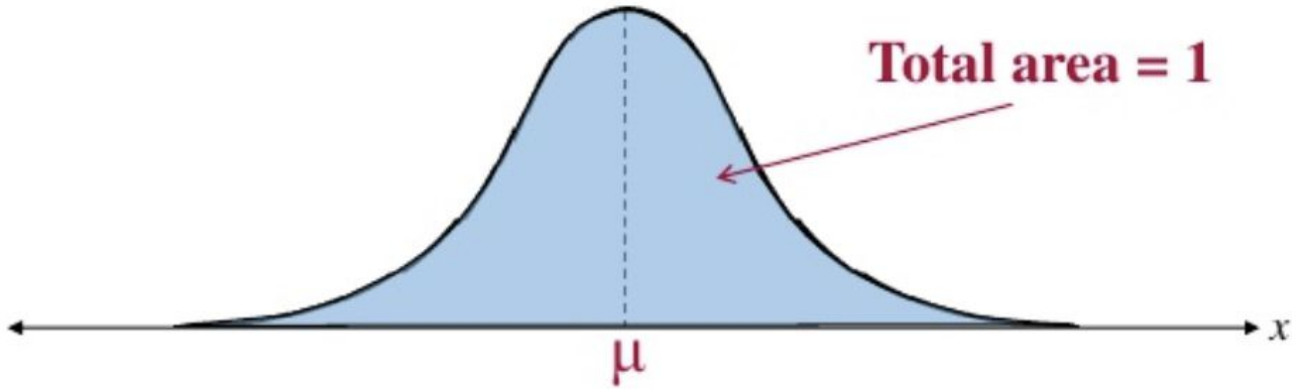
n = 100

p = 0.5

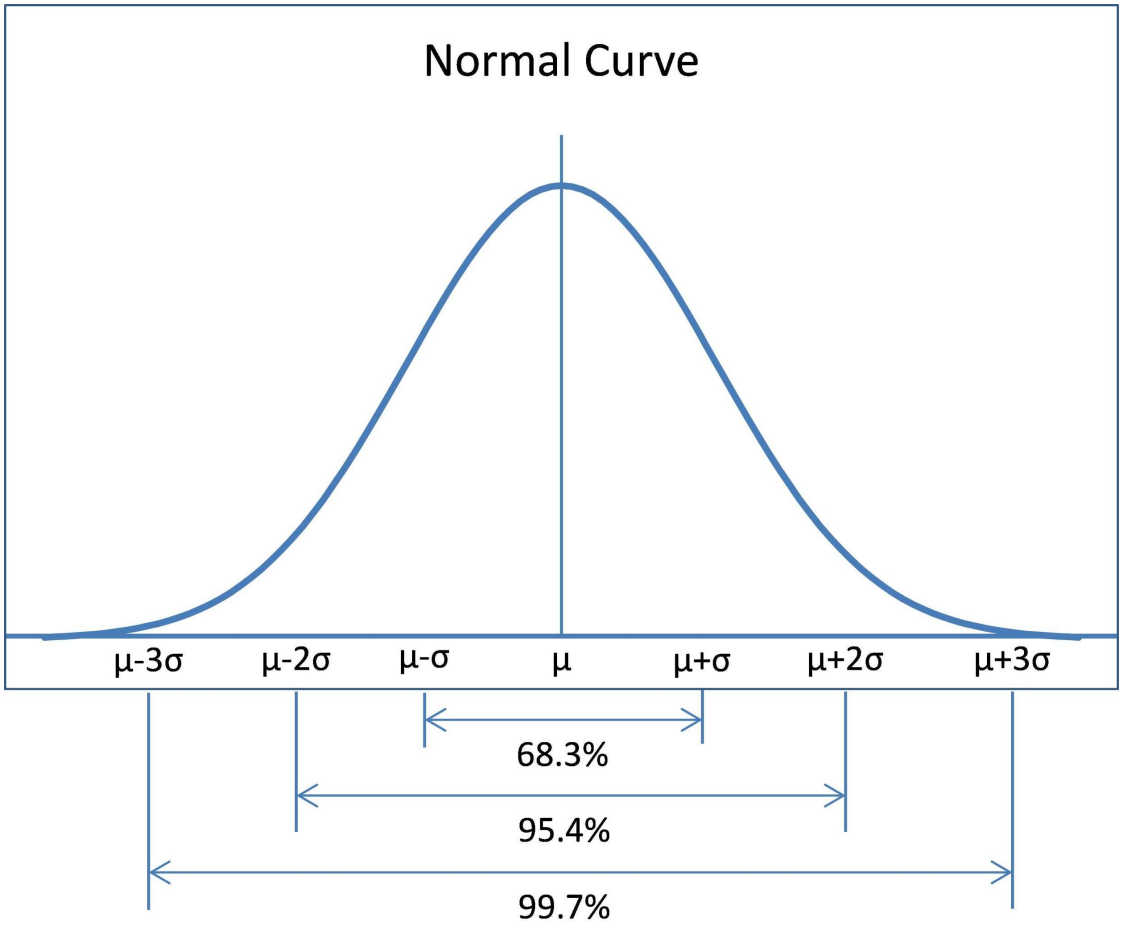
$\sigma = \sqrt{100 * 0.5(1-0.5)} = 5$

# Normal Distribution

- Used for continuous random variable x
- Mean, median and mode are equal
- The curve is bell shaped and symmetrical about the mean
- The total area under the curve = 1
- Each normal distribution has its own mean, $\mu$, and its own standard deviation, $\sigma$, and can be completely determined by these two parameters.



# Area under the normal distribution

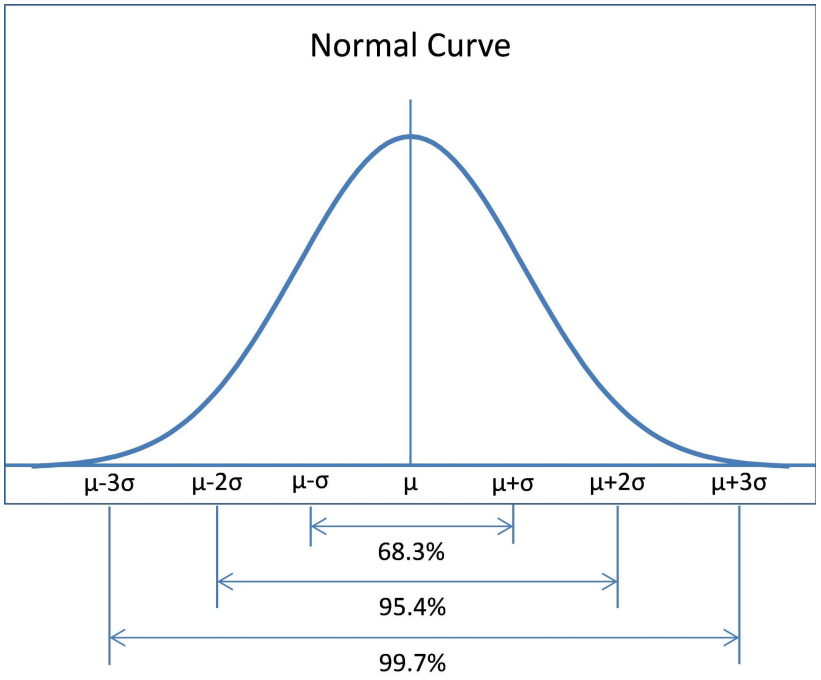


# Normal Distribution

$\mu$ =82

$\sigma$ = 15

- This means 68% scores are between 67 and 97 (1 s.d.)
- 95% scores are between 52 and 112 ( 2 s.d.)




Source: https://www.syncfusion.com/ebooks/statistics/normal-distribution

# Normal Distribution - Example

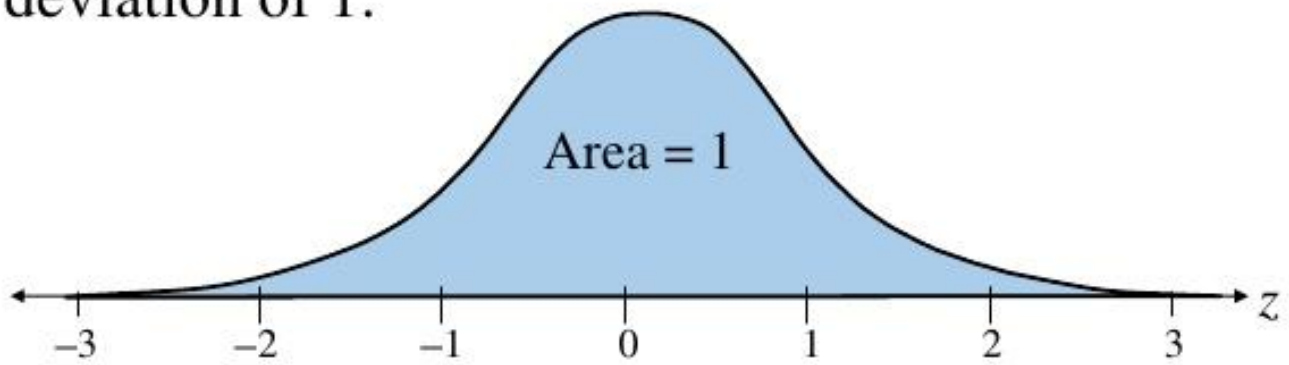
You're a new education director for your country's ministry of education. You have instituted minimum passing scores on the tests that students will take as part of their coursework in mathematics (minimum passing score: 60) and you have a policy that no more than 5% of a class can fail. If this happens, then the entire class must be retested.

One of the classes scored an **average of 82** on their test with a **standard deviation of 15 points**. How do you determine if this class must be retested?

# The Standard Normal Distribution

## Standard normal distribution

- A normal distribution with a mean of 0 and a standard deviation of 1.



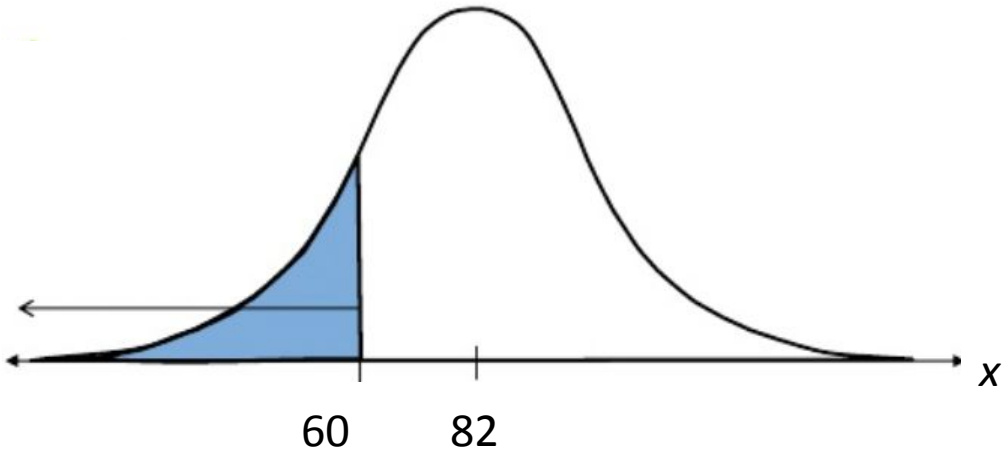

- Any $x$-value can be transformed into a $z$-score by using the formula

$$z = \frac{\text{Value - Mean}}{\text{Standard deviation}} = \frac{x - \mu}{\sigma}$$

# Area under the Normal Distribution

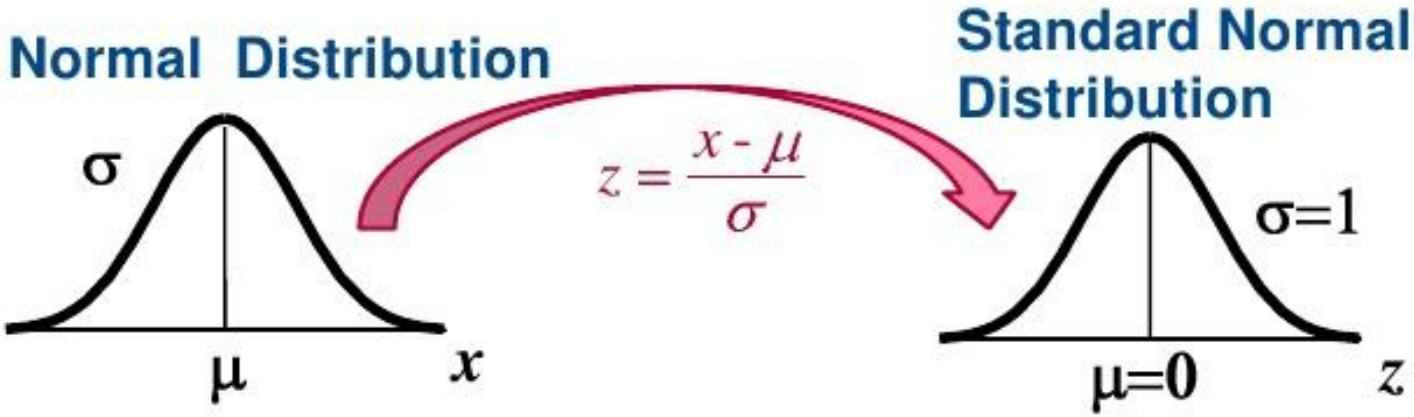
What is the probability that a student enrolled in the class will get a score less than or equal to 60?

$$P(X \leq 60)$$



# The Standard Normal Distribution

- If each data value of a normally distributed random variable $x$ is transformed into a $z$-score, the result will be the standard normal distribution.




- Use the Standard Normal Table to find the cumulative area under the standard normal curve.

# Mathematics test scores example

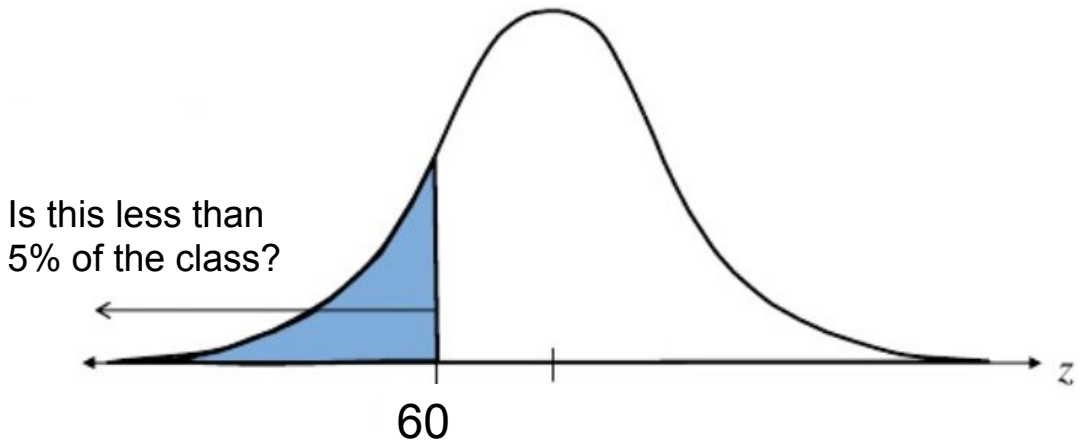
You're a new education director for your country's ministry of education. You have instituted minimum passing scores on the tests that students will take as part of their coursework in mathematics (minimum passing score: 60) and you have a policy that no more than 5% of a class can fail. If this happens, then the entire class must be retested.

One of the classes scored an **average of 82** on their test with a **standard deviation of 15 points**. How do you determine if this class must be retested?

# Mathematics test scores example

We can use standard normal distribution to answer our question!

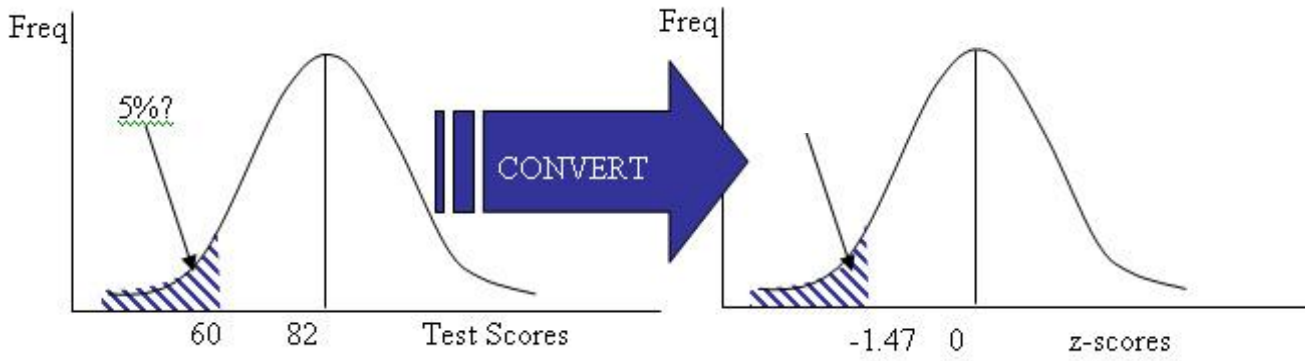
Average class score = 82
Standard deviation = 15
Passing score = 60
Minimum % passing requirement = 5%



# Mathematics test scores example

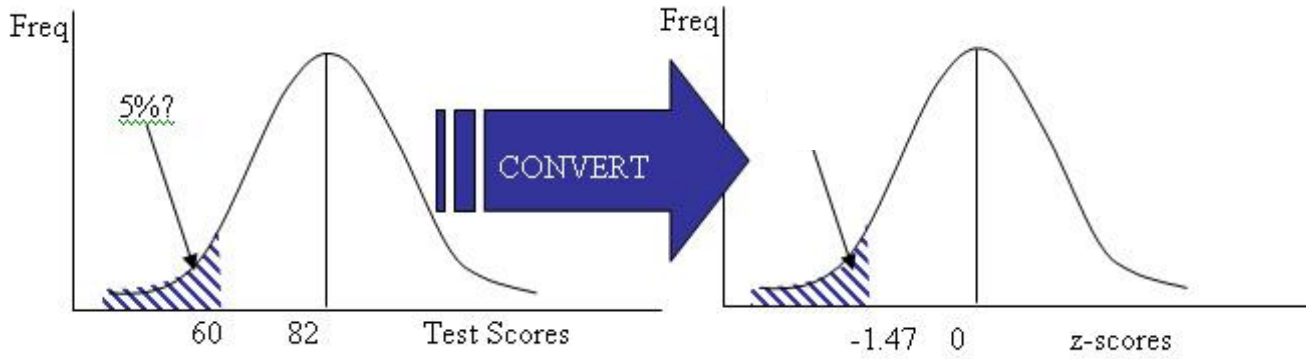
1. Convert the normal distribution curve and to standard normal distribution curve
2. Use the standard normal distribution table to find the respective probability

$$z = \frac{\text{Value - Mean}}{\text{Standard deviation}} = \frac{x - \mu}{\sigma}$$

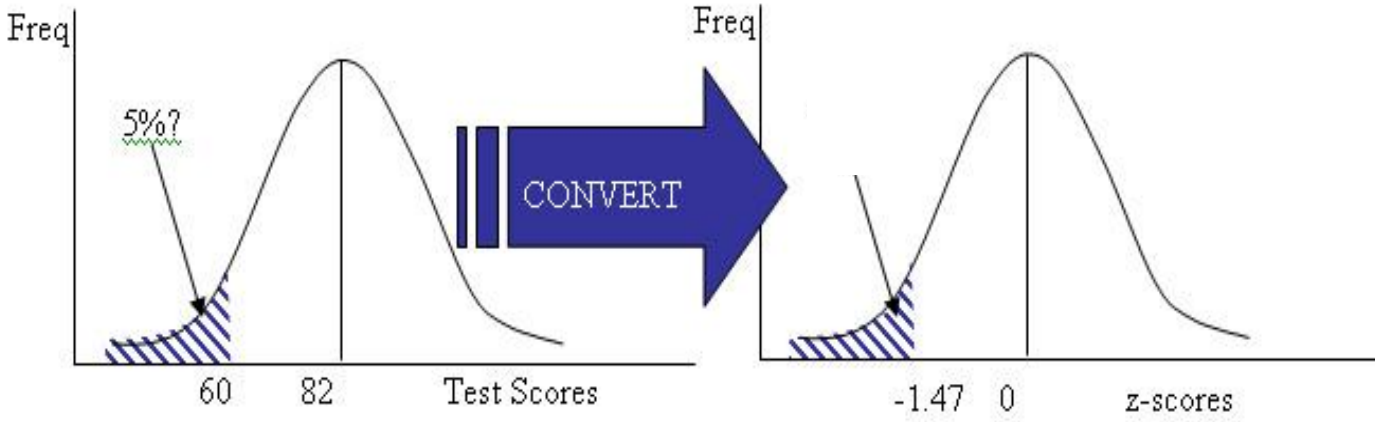


1. Convert the normal distribution curve and to standard normal distribution curve

   What are z-scores for our mean (82) and passing score (60)?




$$z = \frac{60-82}{15} = \frac{-22}{15} = -1.47$$




$$P(X \leq 60) = P(Z \leq -1.47)$$

= Area to the left of -1.47 under the standard normal curve

Z-scores

**TABLE A** Standard Normal cumulative proportions

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| −3.4 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0002 |
| −3.3 | .0005 | .0005 | .0005 | .0004 | .0004 | .0004 | .0004 | .0004 | .0004 | .0003 |
| −3.2 | .0007 | .0007 | .0006 | .0006 | .0006 | .0006 | .0006 | .0005 | .0005 | .0005 |
| −3.1 | .0010 | .0009 | .0009 | .0009 | .0008 | .0008 | .0008 | .0008 | .0007 | .0007 |
| −3.0 | .0013 | .0013 | .0013 | .0012 | .0012 | .0011 | .0011 | .0011 | .0010 | .0010 |
| −2.9 | .0019 | .0018 | .0018 | .0017 | .0016 | .0016 | .0015 | .0015 | .0014 | .0014 |
| −2.8 | .0026 | .0025 | .0024 | .0023 | .0023 | .0022 | .0021 | .0021 | .0020 | .0019 |
| −2.7 | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |
| −2.6 | .0047 | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| −2.5 | .0062 | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| −2.4 | .0082 | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| −2.3 | .0107 | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| −2.2 | .0139 | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| −2.1 | .0179 | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| −2.0 | .0228 | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| −1.9 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| −1.8 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| −1.7 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| −1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| −1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| −1.4 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| −1.3 | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| −1.2 | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| −1.1 | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| −1.0 | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |

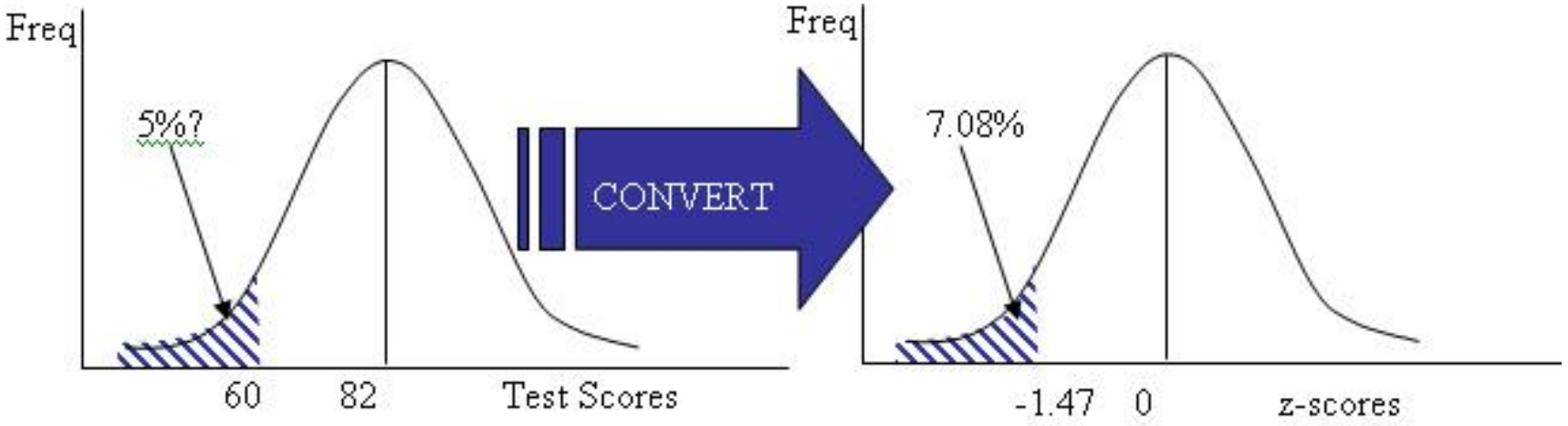
Probabilities

# What is the probability for our Z-score of -1.47?

**TABLE A** Standard Normal cumulative proportions

| z | .00 | .01 | .02 | .03 | .04 | .05 | .06 | .07 | .08 | .09 |
|---|---|---|---|---|---|---|---|---|---|---|
| −3.4 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0002 |
| −3.3 | .0005 | .0005 | .0005 | .0004 | .0004 | .0004 | .0004 | .0004 | .0004 | .0003 |
| −3.2 | .0007 | .0007 | .0006 | .0006 | .0006 | .0006 | .0006 | .0005 | .0005 | .0005 |
| −3.1 | .0010 | .0009 | .0009 | .0009 | .0008 | .0008 | .0008 | .0008 | .0007 | .0007 |
| −3.0 | .0013 | .0013 | .0013 | .0012 | .0012 | .0011 | .0011 | .0011 | .0010 | .0010 |
| −2.9 | .0019 | .0018 | .0018 | .0017 | .0016 | .0016 | .0015 | .0015 | .0014 | .0014 |
| −2.8 | .0026 | | | | | | | .0021 | .0020 | .0019 |
| −2.7 | .0035 | | | | | | | .0028 | .0027 | .0026 |
| −2.6 | .0047 | | | | | | | .0038 | .0037 | .0036 |
| −2.5 | .0062 | | | | | | | .0051 | .0049 | .0048 |
| −2.4 | .0082 | | | | | | | .0068 | .0066 | .0064 |
| −2.3 | .0107 | | | | | | | .0089 | .0087 | .0084 |
| −2.2 | .0139 | | | | | | | .0116 | .0113 | .0110 |
| −2.1 | .0179 | | | | | | | .0150 | .0146 | .0143 |
| −2.0 | .0228 | | | | | | | .0192 | .0188 | .0183 |
| −1.9 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| −1.8 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| −1.7 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| −1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| −1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| −1.4 | | | | | | | | .0708 | .0694 | .0681 |
| −1.3 | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| −1.2 | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| −1.1 | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| −1.0 | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |

Find -1.4 in the left hand column
Move across the row to the column under 0.07
The area to the left of z = -1.47 is 0.0708



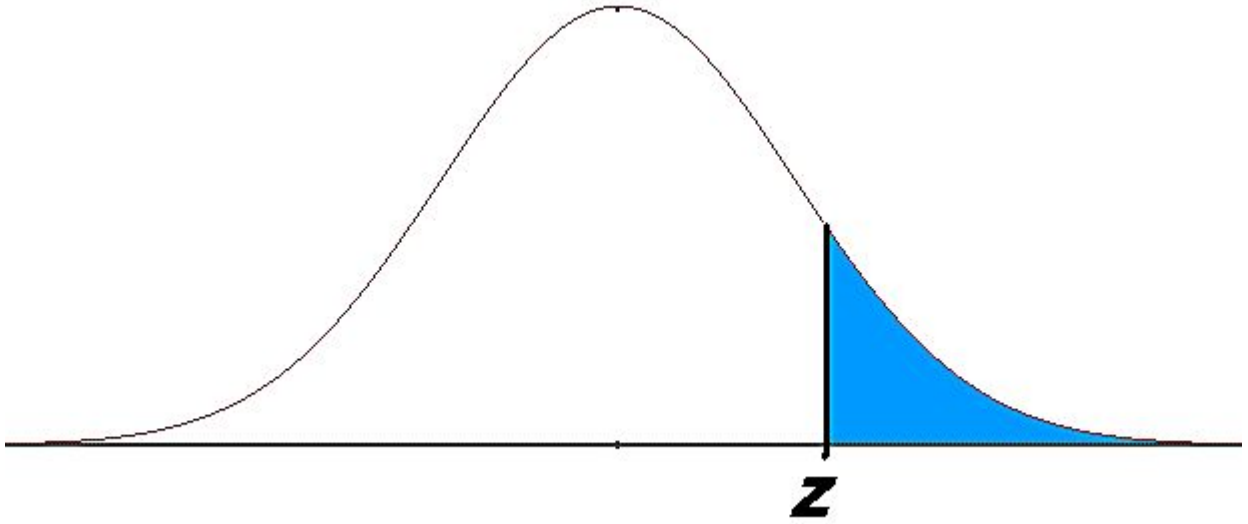

$$P(Z \leq -1.47) = 0.0708$$

Area under the standard normal curve at -1.47 = 0.0708.
Multiply by 100 to get the percentage.
Therefore, 7.08% of the class failed and they sadly have to retest!

# Exercise: Finding area to the right of the normal curve

To find the area to the right of z, use the standard normal table to find the area that corresponds to z and then subtract from 1.

Can you calculate the area to the right of 1.23?

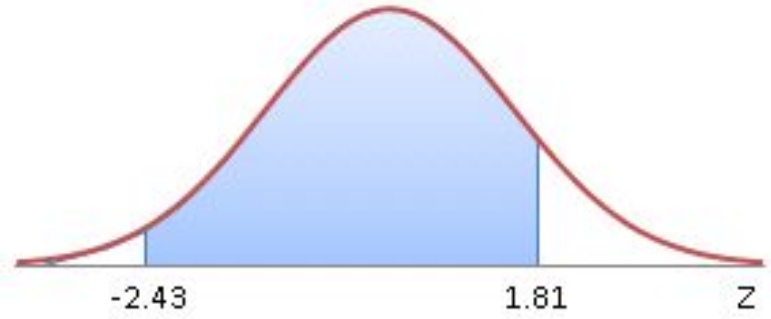


# Exercise: Finding areas under the standard normal curve between two z-scores

To find the area between two z-scores, find the area corresponding to each z-score in the standard normal table. Then subtract the smaller area from the larger area.

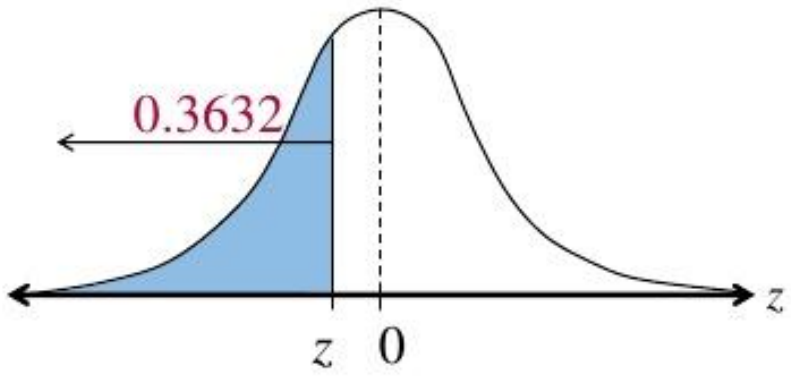
Can you calculate the area between -2.43 and 1.81?

$$P(-2.43 \leq Z \leq 1.81) = ?$$



# Example: Finding a *z*-Score Given an Area

Find the *z*-score that corresponds to a cumulative area of 0.3632.



# Solution: Finding a *z*-Score Given an Area

- Locate 0.3632 in the body of the Standard Normal Table.

| z | .09 | .08 | .07 | .06 | .05 | .04 | .03 |
|---|---|---|---|---|---|---|---|
| −3.4 | .0002 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 |
| −3.3 | .0003 | .0004 | .0004 | .0004 | .0004 | .0004 | .0004 |
| −3.2 | .0005 | .0005 | .0005 | .0006 | .0006 | .0006 | .0006 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| −0.5 | .2776 | .2810 | .2843 | .2877 | .2912 | .2946 | .2981 |
| −0.4 | .3121 | .3156 | .3192 | .3228 | .3264 | .3300 | .3336 |
| −0.3 | .3483 | .3520 | .3557 | .3594 | .3632 | .3669 | .3707 |
| −0.2 | .3859 | .3897 | .3936 | .3974 | .4013 | .4052 | .4090 |
| −0.1 | .4247 | .4286 | .4325 | .4364 | .4404 | .4443 | .4483 |
| −0.0 | .4641 | .4681 | .4721 | .4761 | .4801 | .4840 | .4880 |

The *z*-score is -0.35.

- The values at the beginning of the corresponding row and at the top of the column give the *z*-score.

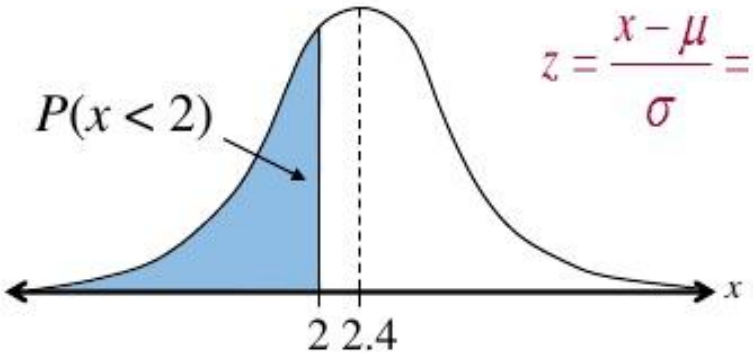
# Example: Finding Probabilities for Normal Distributions

A survey indicates that people use their computers an average of 2.4 years before upgrading to a new machine. The standard deviation is 0.5 year. A computer owner is selected at random. Find the probability that he or she will use it for fewer than 2 years before upgrading. Assume that the variable $x$ is normally distributed.

# Solution: Finding Probabilities for Normal Distributions

Normal Distribution

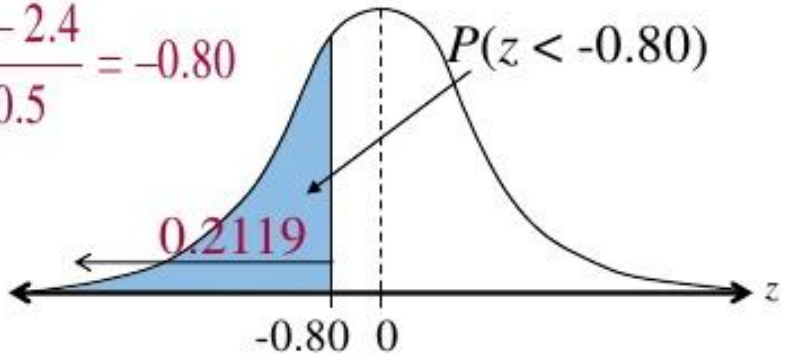
$\mu = 2.4 \quad \sigma = 0.5$




$$z = \frac{x - \mu}{\sigma} = \frac{2 - 2.4}{0.5} = -0.80$$

Standard Normal Distribution

$\mu = 0 \quad \sigma = 1$




$$P(x < 2) = P(z < -0.80) = \mathbf{0.2119}$$

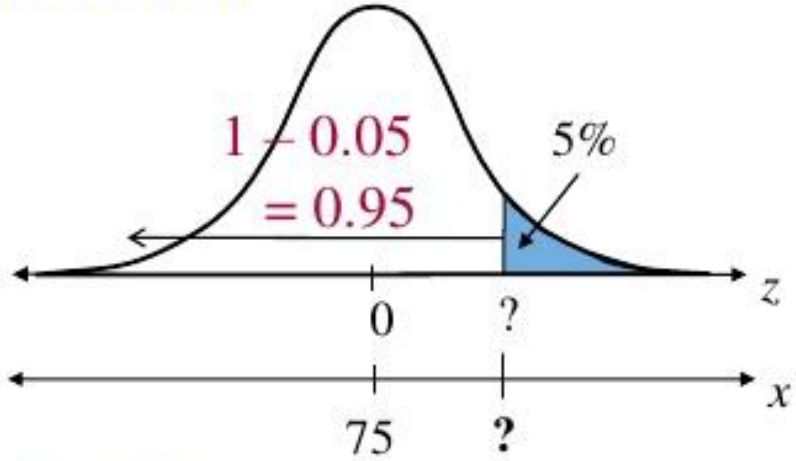
## Example: Finding a Specific Data Value

Scores for a civil service exam are normally distributed, with a mean of 75 and a standard deviation of 6.5. To be eligible for civil service employment, you must score in the top 5%. What is the lowest score you can earn and still be eligible for employment?

# Example: Finding a Specific Data Value

Scores for a civil service exam are normally distributed, with a mean of 75 and a standard deviation of 6.5. To be eligible for civil service employment, you must score in the top 5%. What is the lowest score you can earn and still be eligible for employment?

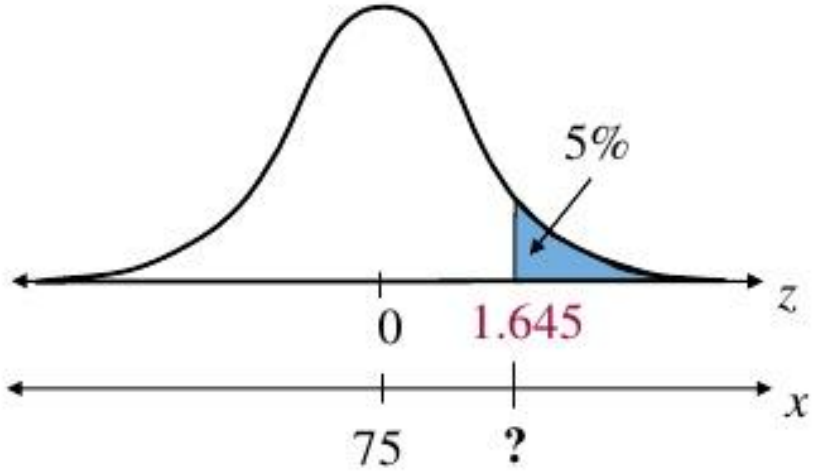
**Solution:**




An exam score in the top 5% is any score above the 95th percentile. Find the $z$-score that corresponds to a cumulative area of 0.95.

# Solution: Finding a Specific Data Value

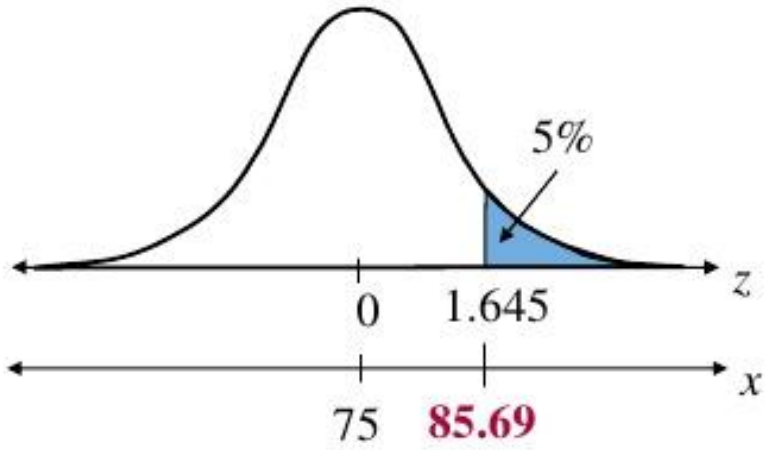
From the Standard Normal Table, the areas closest to 0.95 are 0.9495 ($z = 1.64$) and 0.9505 ($z = 1.65$). Because 0.95 is halfway between the two areas in the table, use the $z$-score that is halfway between 1.64 and 1.65. That is, $\mathbf{z = 1.645}$.



# Solution: Finding a Specific Data Value

Using the equation $x = \mu + z\sigma$

$$x = 75 + 1.645(6.5) \approx 85.69$$




The lowest score you can earn and still be eligible for employment is 86.

# Module 3: Univariate Analysis - Confidence Interval

# Statistical Inference

You work for the Ministry of Poverty Alleviation, and the Minister, would like to understand the impact of a new welfare program on the income of beneficiaries. But the welfare program in question was implemented in the whole country. The Minister needs an answer in about an hour. What can you do? Can you simply look at a sample of beneficiaries, determine the average impact on income, and use this to help answer the question?

# Statistical Inference

- The process of drawing conclusions about a population on the basis of sample data

- When sampling to make inferences, use a simple random variable to eliminate bias and make accurate predictions

- **parameter** ~ describes a population. Example, the mean income of households in Nepal
- **statistic** ~ computed from the sample. Example, the mean income of a random sample of households in Nepal

- Use a statistic to estimate an unknown parameter.

- $\boldsymbol{\mu}$ ~ population mean, $\boldsymbol{\sigma}$ ~ population standard deviation
  $\bar{x}$ ~ sample mean, $\boldsymbol{s}$~ sample standard deviation

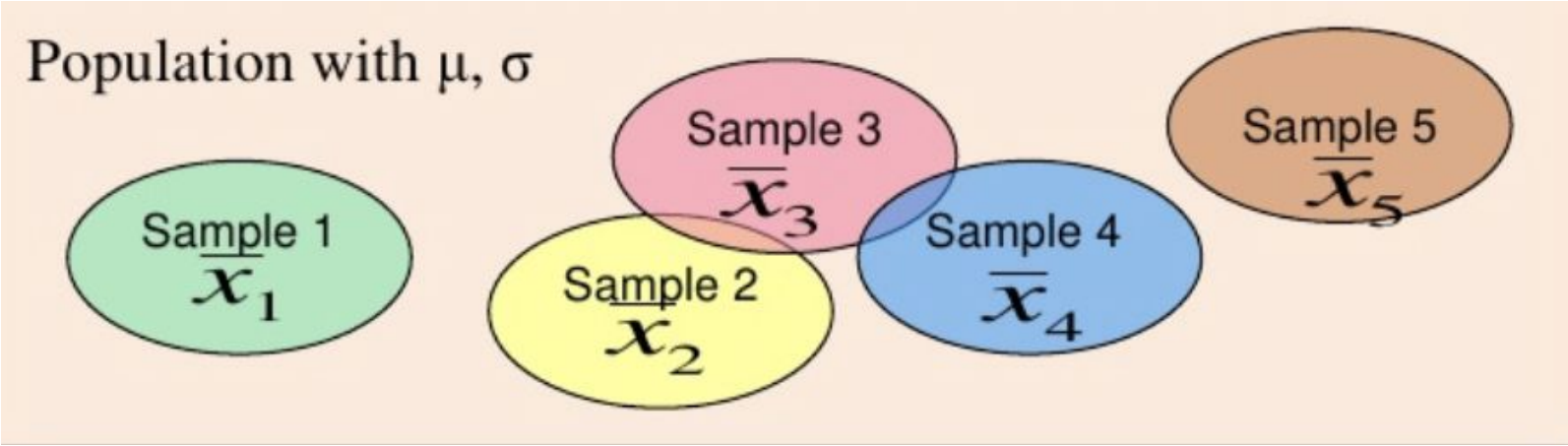
# Types of Statistical Inference

- Confidence Interval - used to estimate a population parameter from sample statistics

- Tests of significance - used to assess the evidence provided by data about some claim concerning a population parameter.

# Statistical Inference

- Randomly sample 100 individuals and get their income.
- Find the mean and standard deviations of income of those samples.
- Keep doing this over and over again, 1000 times.
- Let's collect all the 1000 means and draw a histogram of all these means.
- The histogram represents the distribution of all possible sample means $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_{1000}$




This distribution is called the **sampling distribution of the mean OR the sampling distribution of the sample means.**

# Properties of Sampling Distributions of Sample Means

1. The mean of the sample means, $\mu_{\bar{x}}$, is equal to the population mean $\mu$.

$$\mu_{\bar{x}} = \mu$$

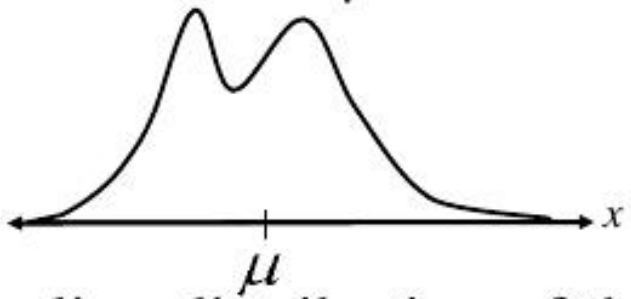
2. The standard deviation of the sample means, $\sigma_{\bar{x}}$, is equal to the population standard deviation, $\sigma$ divided by the square root of the sample size, $n$.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

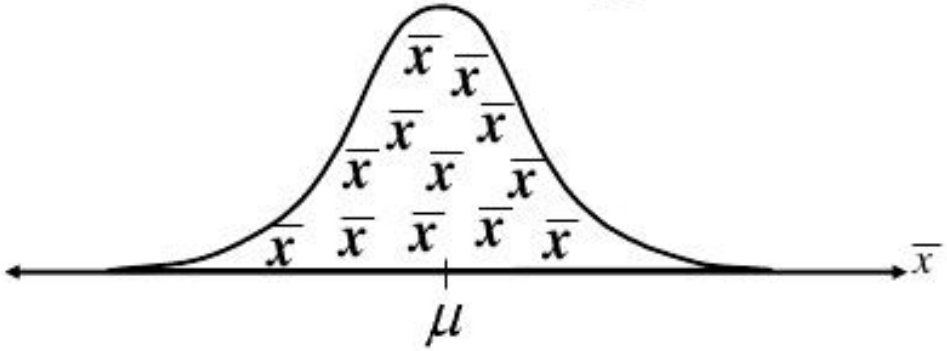
   - Called the **standard error of the mean**.

# The Central Limit Theorem

1. If samples of size $n \geq 30$, are drawn from any population with mean = $\mu$ and standard deviation = $\sigma$,

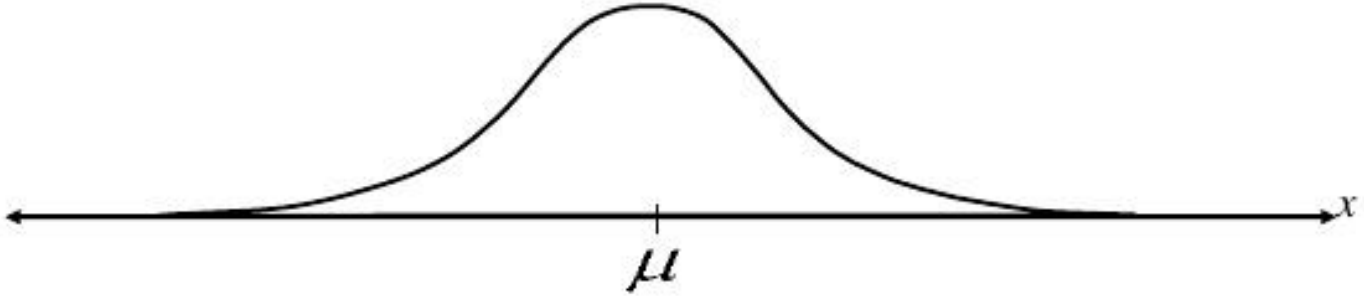



then the sampling distribution of the sample means approximates a normal distribution. The greater the sample size, the better the approximation.

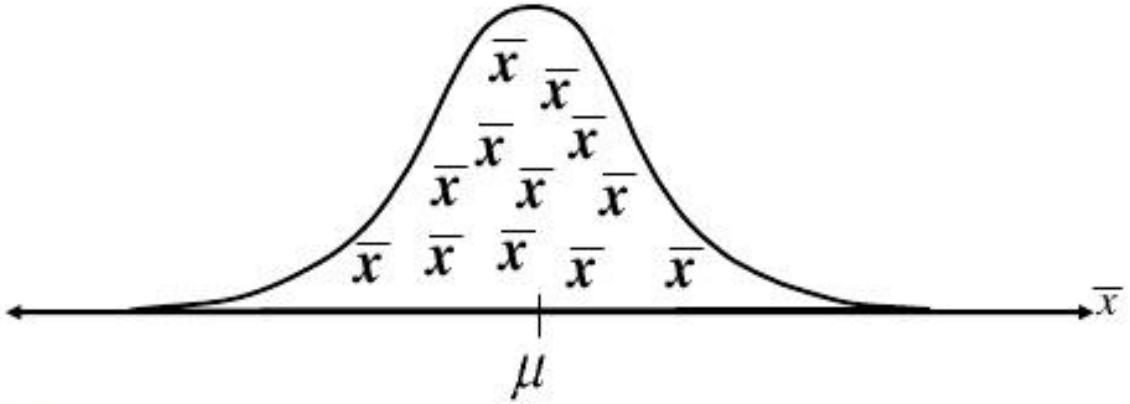


# The Central Limit Theorem

2. If the population itself is normally distributed,

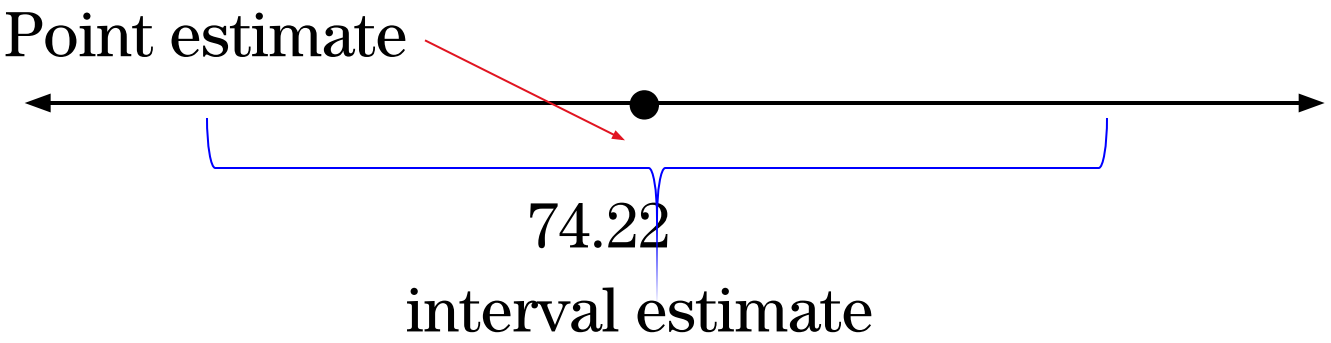



the sampling distribution of the sample means is normally distribution for ***any*** sample size $n$.



# Point Estimate for Population $\mu$

- **point estimate** - single value estimate for a population parameter
- point estimate of the population mean, $\mu$, is the ***sample*** mean, $\bar{x}$.




The point estimate for the population mean ~ 74.22.

- **interval estimate** - an interval, or range of values to estimate a population parameter.

**How confident do we want to be that the interval estimate contains the population mean, $\mu$?**

# Confidence Interval

- Estimates a population parameter by using statistics of a sample.
- Example, estimate the average income (parameter) based on the average income from a random sample of 100 individuals (statistic).
- Because sample results will vary, add a measure of that variability to the estimate, called the margin of error (MOE)

**Confidence Interval = Sample Statistic ± MOE**

# Confidence Interval

- Ultimate goal when making an estimate using a confidence interval: a small margin of error (MOE) The narrower the interval, the more precise the results!

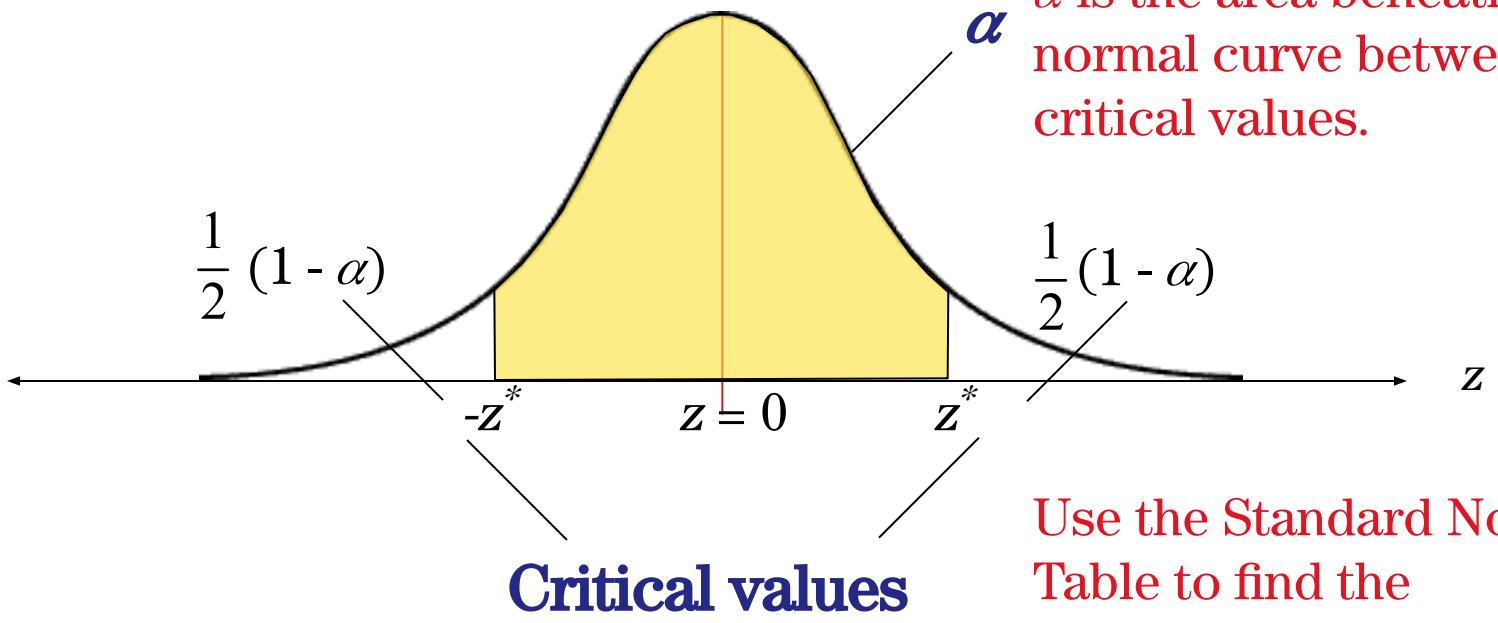
- Three factors affect the size/width of the MOE
  - ✓ The confidence level, $\alpha$
  - ✓ The sample size, $n$
  - ✓ The amount of variability in the population, $\sigma$

# Confidence Level $\alpha$

- percentage of the time the result would be correct/fall under the interval if we took numerous random samples
- probability that the interval will capture the true parameter value in repeated samples
- Typical confidence levels ~ 90%, 95% or 99%

# Confidence Level $\alpha$

The **level of confidence $\alpha$** is the probability that the interval estimate contains the population parameter.



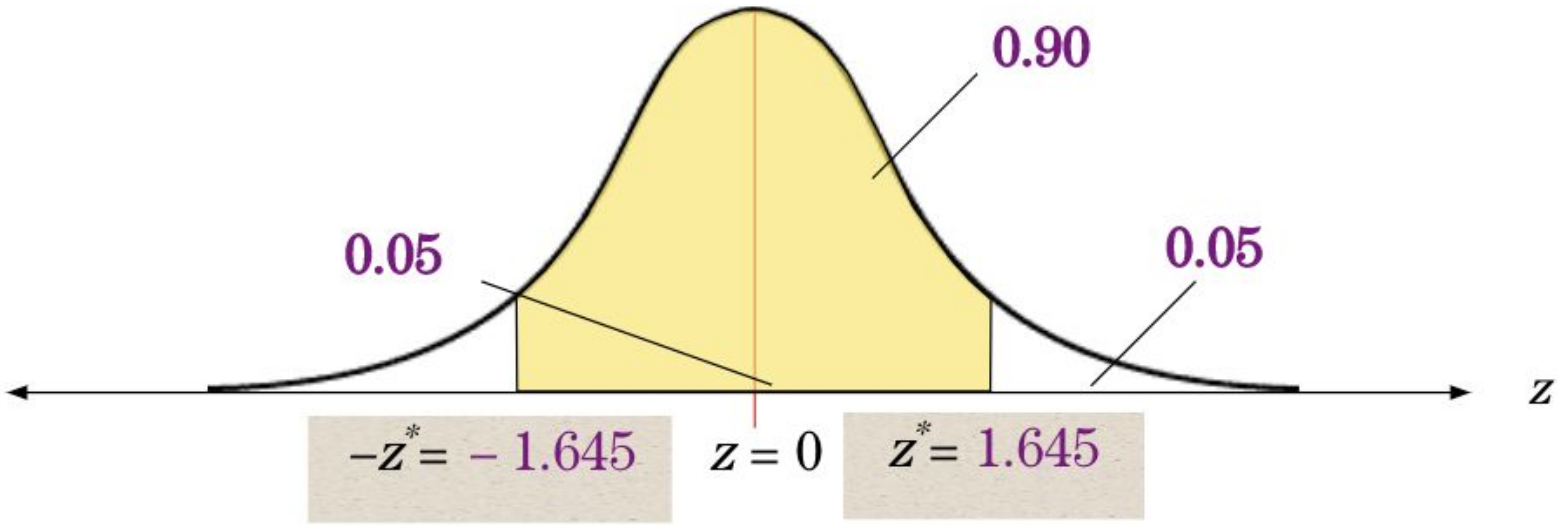

$\alpha$ is the area beneath the normal curve between the critical values.

Use the Standard Normal Table to find the corresponding *z*-scores.

The remaining areas in the tails is 1-$\alpha$.

# Common Confidence Levels $\alpha$

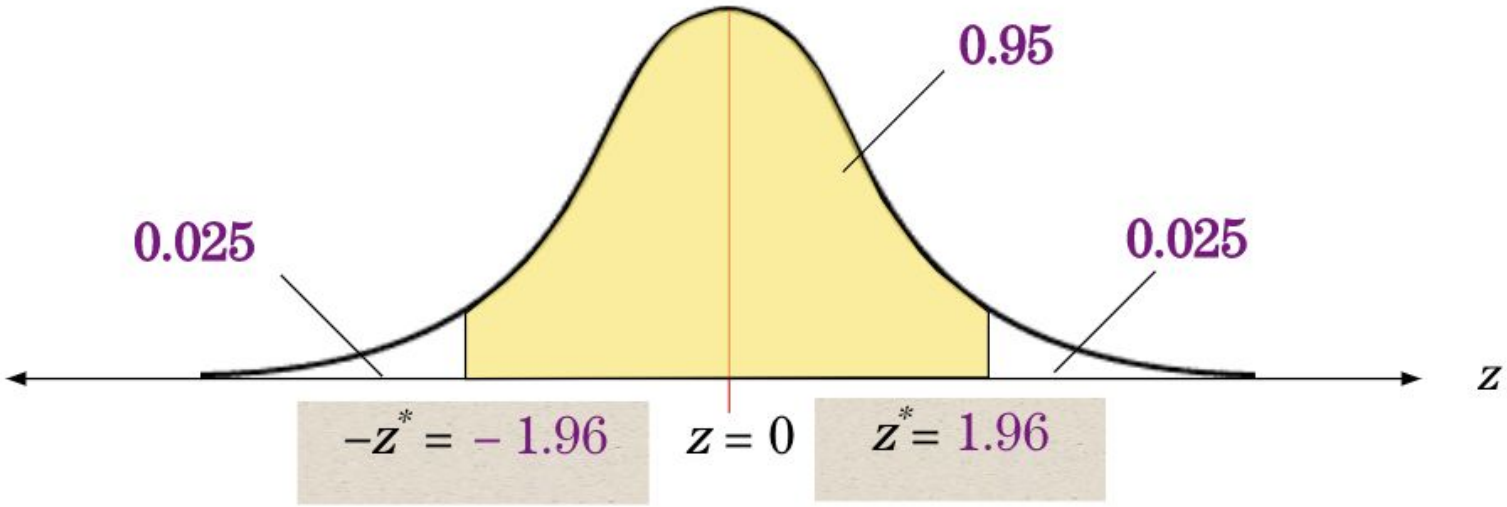
If the level of confidence is 90%, this means that we are 90% confident that the interval contains the population mean, $\mu$.




The corresponding $z$-scores are $\pm$ 1.645.

# Common Confidence Levels $\alpha$

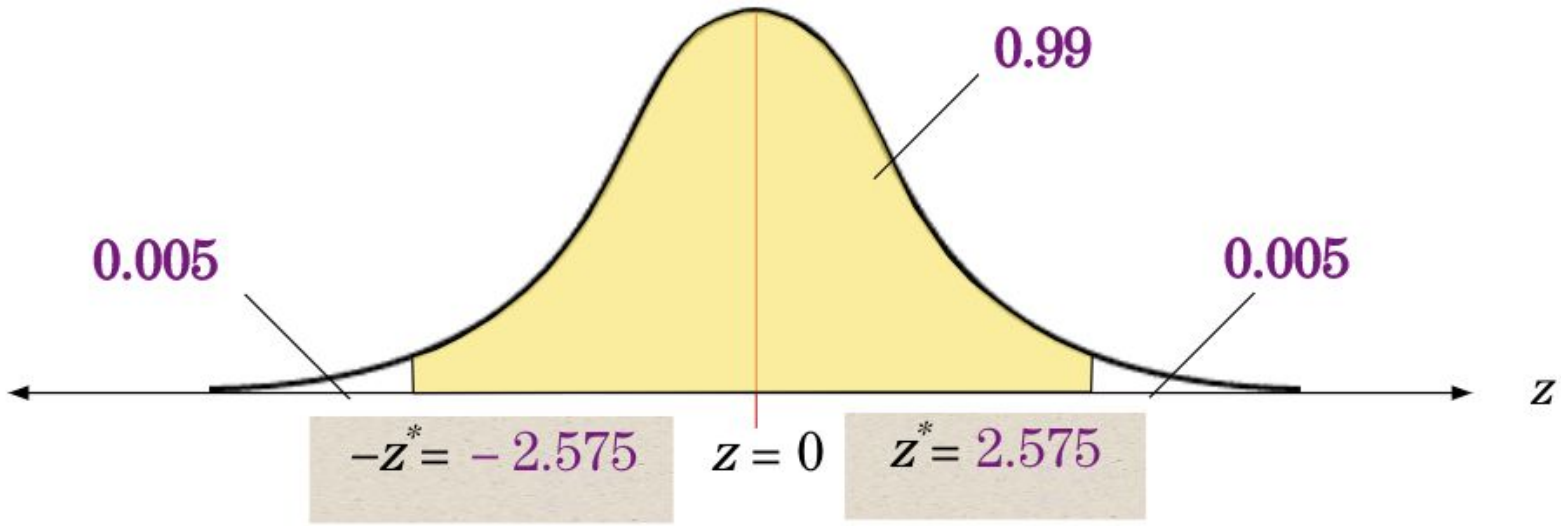
If the level of confidence is 95%, this means that we are 95% confident that the interval contains the population mean, $\mu$.




The corresponding $z$-scores are $\pm$ 1.96.

# Common Confidence Levels $\alpha$

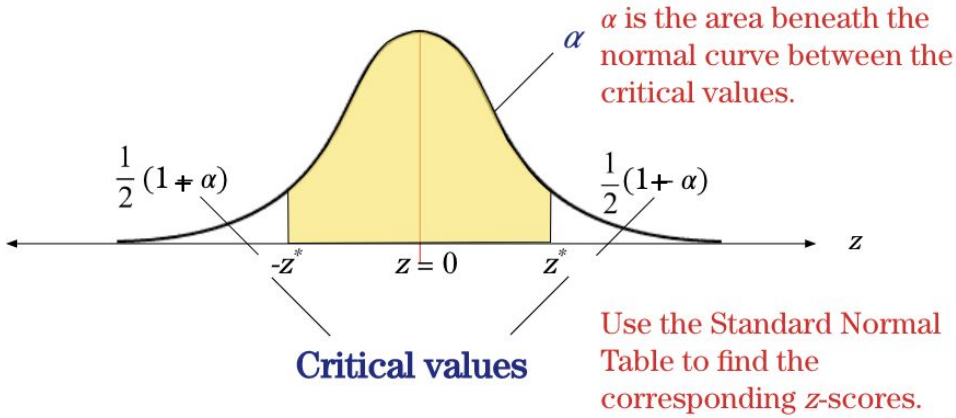
If the level of confidence is 99%, this means that we are 99% confident that the interval contains the population mean, $\mu$.




The corresponding $z$-scores are $\pm$ 2.575.

# Confidence Interval for a population mean

The **level of confidence $\alpha$** is the probability that the interval estimate contains the population parameter.




**Confidence Interval for the mean of a population =**

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

$$\text{CI} = \left[\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right]$$

$$z^* = z_{\frac{1}{2}+\frac{\alpha}{2}} = z_{\frac{1+\alpha}{2}}$$

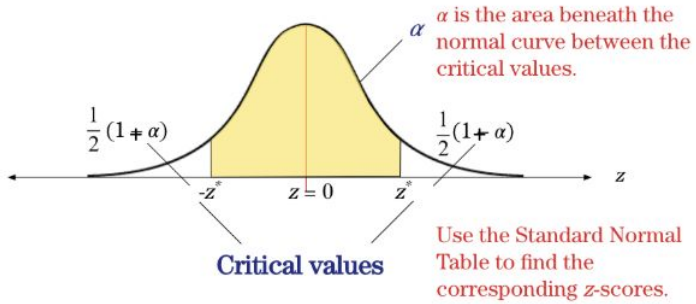
# Margin of Error

$$\text{MOE} = z^* \frac{\sigma}{\sqrt{n}}$$

$$\text{CI} = \left[\bar{x} - z^* \frac{\sigma}{\sqrt{n}}, \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right]$$

$$\bar{x} - \text{MOE} \leq \mu \leq \bar{x} + \text{MOE}$$

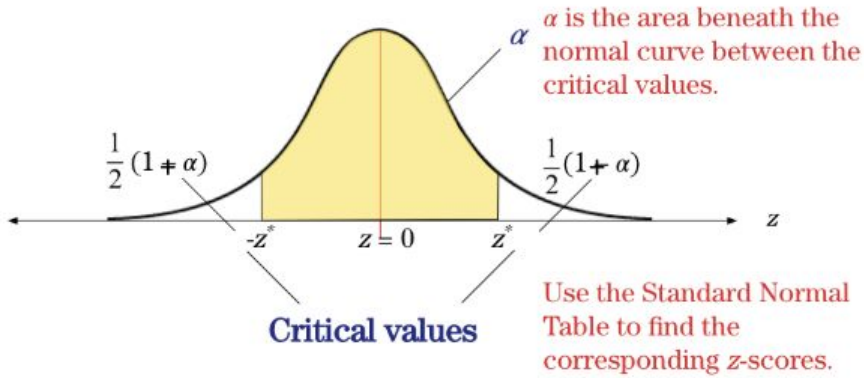
The **level of confidence** $\alpha$ is the probability that the interval estimate contains the population parameter.




The difference between the point estimate and the actual population parameter value

# Confidence Interval for a population mean

The **level of confidence** $\alpha$ is the probability that the interval estimate contains the population parameter.




**Margin of Error**

$$\text{MOE} = z^* \frac{\sigma}{\sqrt{n}}$$

$$z^* = z_{\frac{1}{2}+\frac{\alpha}{2}} = z_{\frac{1+\alpha}{2}}$$

**sample size n to achieve a certain MOE**

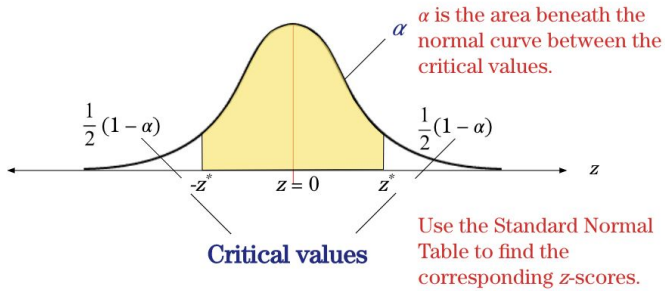
$$n = \left( \frac{z^* \sigma}{MOE} \right)^2$$

# Lab for Confidence Interval

**You measure the weights of a random sample of 61 males runners in a community. The respective sample mean and standard deviation are 60 kg and 5 kg. Compute a 99% confidence interval for $\mu$ assuming that the weights of male runners follow a Normal distribution.**

# Lab for Confidence Interval

The **level of confidence** $\alpha$ is the probability that the interval estimate contains the population parameter.




Solution: $\bar{x} = 60$, $\sigma = 5$, $n = 61$, $\alpha = 0.99$. So,

$$\frac{1+\alpha}{2} = \frac{1+0.99}{2} = 0.995, \quad z^* = z_{0.995}$$

From the standard normal table, $z_{0.995} = 2.58$. Hence, the confidence interval is

$$\left[\bar{x} - z_{0.995}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.995}\frac{\sigma}{\sqrt{n}}\right]$$

$$= \left[60 - 2.58\frac{5}{\sqrt{61}}, 60 + 2.58\frac{5}{\sqrt{61}}\right]$$

$$= [58.35, 61.65]$$

$$MOE = 2.58\frac{5}{\sqrt{61}} = 1.65$$

# Confidence Interval Lab

- Confidence interval on paper
- Confidence interval on Python

Google Doc for “confidence interval” classwork: https://docs.google.com/document/d/1O59LdInzG8VbONHh9nP_LXZ55evl3o6K5C6wW6RWIOk/edit?usp=sharing

# Notebook for “Confidence Interval” Lab

https://colab.research.google.com/drive/1_2LbRbG7K9bGQptV4DxBSa15KexJCs4c

# Application: Spam Detection

- **Spam detection**, a document classification task
- Uses **Naive Bayes Classifier**
- Naive Bayes Classifier is based on "**Bayes Theorem**"
- "Bayes Theorem" is extension of **conditional probability**

**Colab Link:**

***https://colab.research.google.com/drive/1Do643AbxFsK1sp9NgkNPZHhfmq6mHRA_***

# Module 4: Univariate Analysis - Hypothesis Testing

# Hypothesis Testing

Process that uses sample statistics to test a claim about the value of a population parameter

- Making an inference or decision about a parameter, eg, population mean ($\mu$), for instance, asking if it has changed based on a sample of data
- Testing a prior belief about the value of a population parameter

EXAMPLE:
A non profit organization claims that 25% of all women in Nepal are literate. This is a claim about the proportion (parameter) of all women in Nepal (population) who are literate.

# Hypothesis Testing

A non profit organization claims that 25% of all women in Nepal are literate. This is a claim about the proportion (parameter) of all women in Nepal (population) who are literate.

Claim:
the population proportion (p) of literate women in Nepal = 0.25.
This claim is the null hypothesis (denoted $H_0$).

If you're out to test this claim or you're questioning the claim:

- the actual proportion of women who are literate is lower than 0.25, based on your observations or survey, or,
- the proportion may be higher than 0.25, or,
- "No, the proportion is not 0.25."

This hypothesis is the alternative hypothesis (denoted $H_A$).

# Stating a Hypothesis

A **null hypothesis $H_0$** := hypothesis about population that contains a statement of equality such as ≤, =, or ≥.
says nothing new is happening

A **alternative hypothesis** $H_A$ := the complement of the null hypothesis. It is a statement that must be true if $H_0$ is false and contains a statement of inequality such as >, ≠, or <.

To write the null and alternative hypotheses, translate the claim made about the population parameter from a verbal statement to a mathematical statement.

# Hypothesis Testing

A non profit organization claims that 25% of all women in Nepal are literate. This is a claim about the proportion (parameter) of all women in Nepal (population) who are literate.

- $H_0 : p_{literate} = 0.25$
- $H_A : p_{literate} \neq 0.25$ or, $H_A : p_{literate} < 0.25$ or, $H_A : p_{literate} > 0.25$

# Steps for Hypothesis Testing

- Set up $H_0$ and $H_A$
- Always begin the hypothesis test assuming that the null hypothesis is true.
- **Test the claim using sample data**
- At the end of the test, one of two decisions will be made:
  1. reject the null hypothesis if you have sufficient evidence against $H_0$, or
  2. fail to reject the null hypothesis since there is no sufficient evidence to support it

# Testing a claim

1. Level of Significance
2. Test Statistic
3. P-values
4. Compare p-values with the level of significance

# Level of Significance

In a hypothesis test, the **level of significance** is your maximum allowable probability of making an error on rejecting $H_0$. It is denoted by the lowercase Greek letter $\alpha$.

Commonly used levels of significance:

$\alpha = 0.10$ $\alpha = 0.05$ $\alpha = 0.01$

# Test Statistic

After stating the null and alternative hypotheses and specifying the level of significance, a random sample is taken from the population and sample statistics are calculated.

The statistic that is compared with the parameter in the null hypothesis is called the **test statistic**.

| Population parameter | Test statistic | Standardized test statistic |
|---|---|---|
| $\mu$ | $\overline{x}$ | $z$ $(n \geq 30)$ |
| $p$ | $\hat{p}$ | $z$ |

$$z_{value} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

# *P*-values

- ***P*-value** (or **probability value**) measures how likely it was that you would have gotten your sample results if the null hypothesis were true.
- The smaller the P-value, the stronger the evidence against $H_0$ provided by the data

There are three types of hypothesis tests – a left-, right-, or two-tailed test. The type of test depends on the region of the sampling distribution that favors a rejection of $H_0$. This region is indicated by the alternative hypothesis.

# How to find *P*-values?

To find the p-value for your test statistic:

1. Look up the location of your test statistic on the standard normal distribution from the table.

$$z_{\text{value}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

2. Find the percentage chance of being at or beyond that value in the same direction:

$H_A: \mu < k \quad P(Z \leq z_{\text{value}})$

$H_A: \mu > k \quad P(Z \geq z_{\text{value}})$

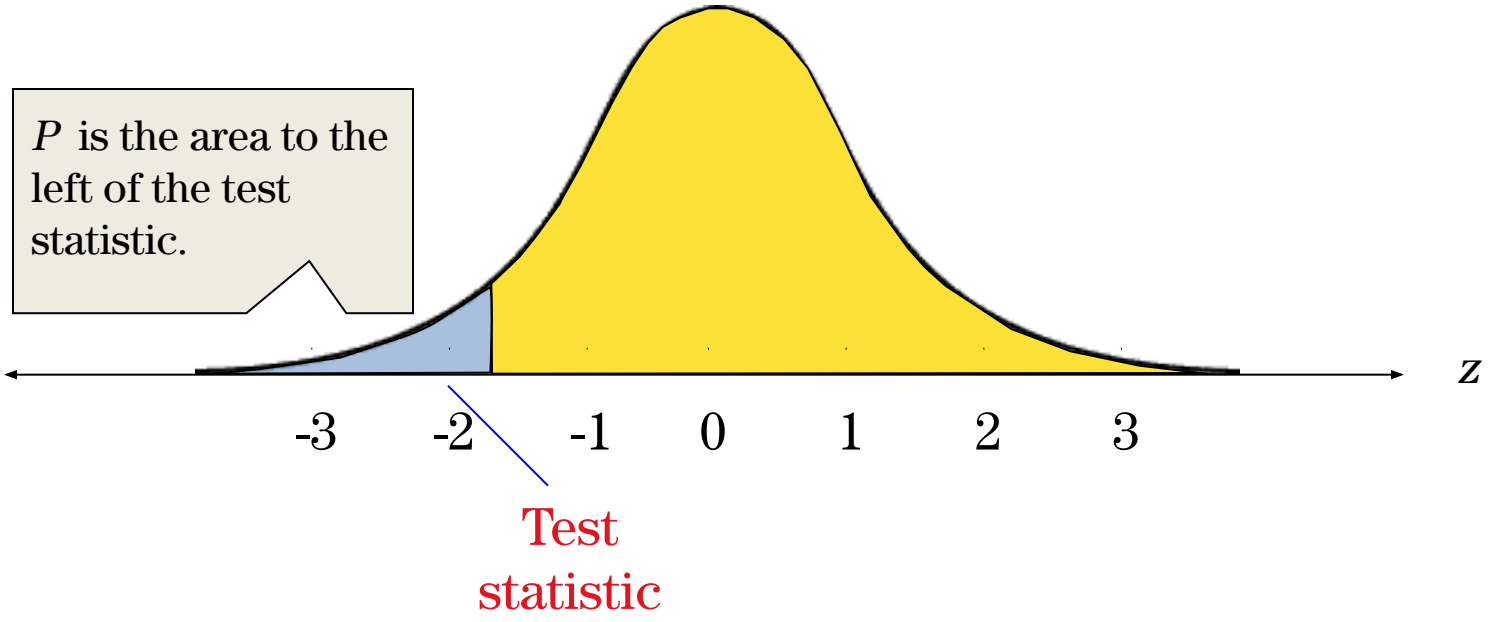
$H_A: \mu \neq k \quad 2 * P(Z \leq z_{\text{value}})$

# Left-tailed Test

1. If the alternative hypothesis contains the less-than inequality symbol (<), the hypothesis test is a **left-tailed test**.

$H_0: \mu \geq k$

$H_A: \mu < k$

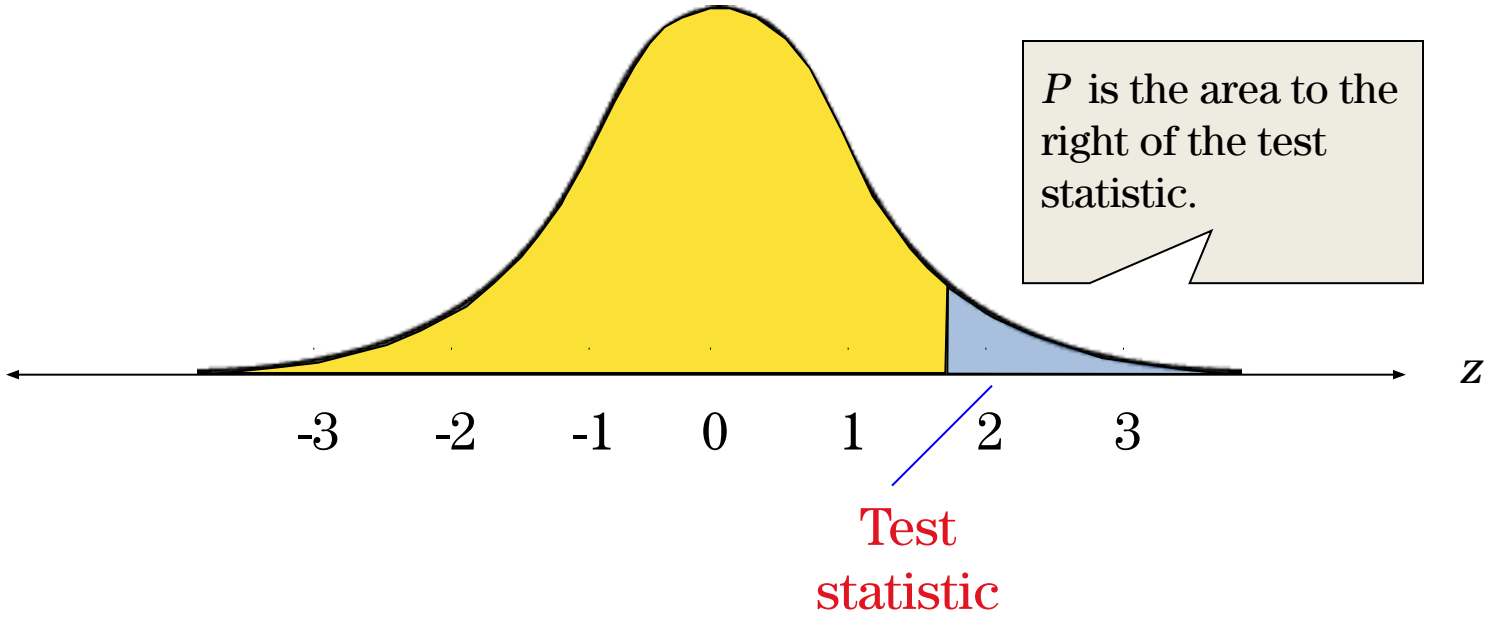


# Right-tailed Test

2. If the alternative hypothesis contains the greater-than symbol (>), the hypothesis test is a **right-tailed test**.

$H_0: \mu \leq k$

$H_A: \mu > k$

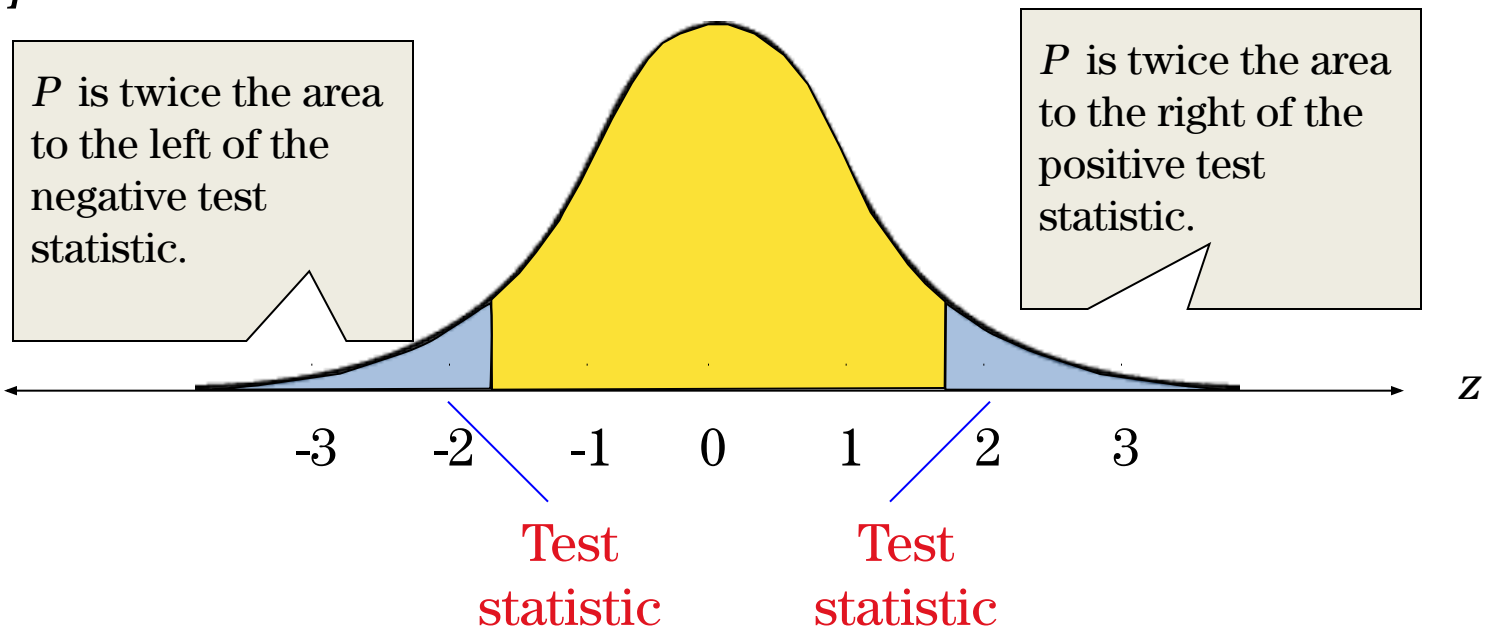


# Two-tailed Test

3. If the alternative hypothesis contains the not-equal-to symbol (≠), the hypothesis test is a **two-tailed test**. In a two-tailed test, each tail has an area of $\frac{1}{2}P$.

$H_0: \mu = k$

$H_A: \mu \neq k$

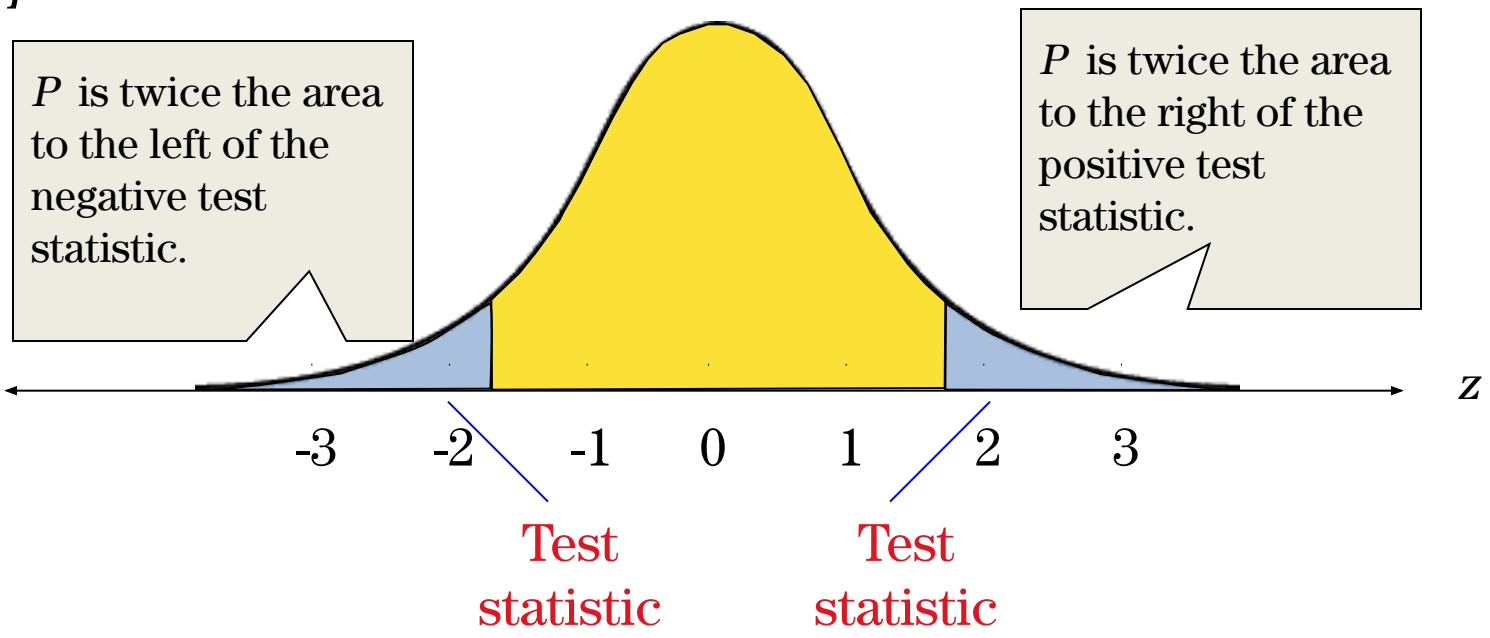


# Two-tailed Test

3. If the alternative hypothesis contains the not-equal-to symbol (≠), the hypothesis test is a **two-tailed test**. In a two-tailed test, each tail has an area of $\frac{1}{2}P$.

$H_0: \mu = k$

$H_A: \mu \neq k$



# Making a Decision

**Decision Rule Based on *P*-value**

To use a *P*-value to make a conclusion in a hypothesis test, compare the *P*-value with $\alpha$.

1. If *P*-value $\leq \alpha$, then reject $H_0$.
2. If *P*-value $> \alpha$, then fail to reject $H_0$.

| | Claim | |
|---|---|---|
| **Decision** | **Claim is $H_0$** | **Claim is $H_A$** |
| Reject $H_0$ | There is enough evidence to reject the claim. | There is enough evidence to support the claim. |
| Do not reject $H_0$ | There is not enough evidence to reject the claim. | There is not enough evidence to support the claim. |

# Interpreting a Decision

**Example**:
You perform a hypothesis test for the following claim. How should you interpret your decision if you reject $H_0$? If you fail to reject $H_0$?

$H_0$: (Claim) A non profit organization claims that 25% of all women in Nepal are literate.

If $H_0$ is rejected, you should conclude "there is sufficient evidence to indicate that the organization's claim is false."

If you fail to reject $H_0$, you should conclude "there is *not* sufficient evidence to indicate that the organization's claim is false."

# Steps for Hypothesis Testing

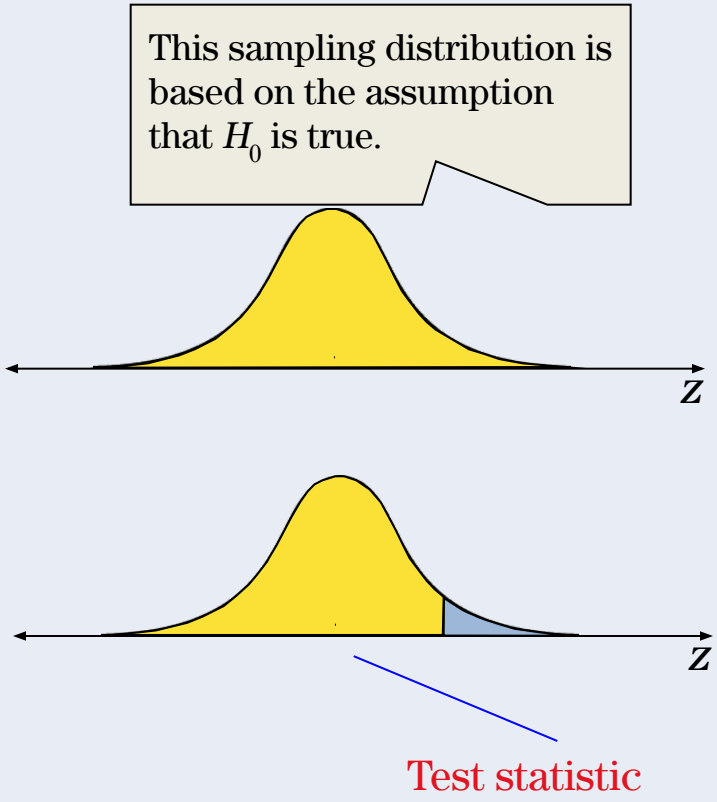
- State the claim mathematically and verbally. Identify the null and alternative hypotheses.

  $H_0$: ?  $H_A$: ?

- Specify the level of significance.

  $\alpha$ = ?

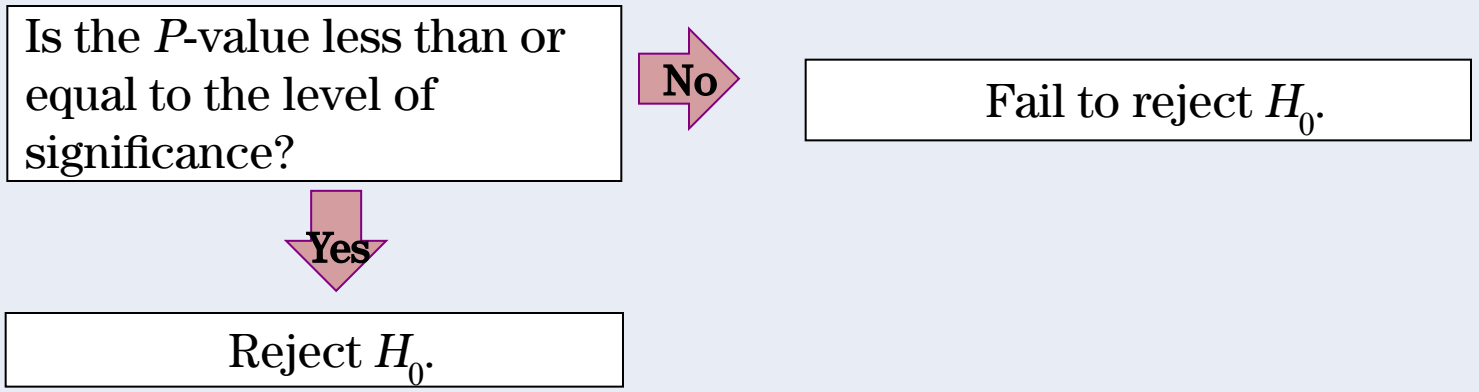
- Calculate the test statistic and its standardized value.




Continued.

# Steps for Hypothesis Testing

- Find the *P*-value.
- Use the following decision rule.




- Write a statement to interpret the decision in the context of the original claim.

These steps apply to left-tailed, right-tailed, and two-tailed tests.

# Hypothesis Testing -- Example

The UN has claimed that its Millennium Project has substantially improved the average GDP/Capita for non-G8 countries. Non-G8 countries typically have an average per capita GDP of $1,250 or less.

A random sample of 49 non-G8 countries revealed an average GDP/capita of $1,400 with a standard deviation of $700. Did the Millennium Project increase average GDP/capita for non-G8 countries? Assume that given the level of significance is 0.1.

# Hypothesis Testing

The UN has claimed that its Millennium Project has substantially improved the average GDP/Capita for non-G8 countries. Non-G8 countries typically have an average per capita GDP of $1,250 or less.

A random sample of 49 non-G8 countries revealed an average GDP/capita of $1,400 with a standard deviation of $700. Did the Millennium Project increase average GDP/capita for non-G8 countries? Assume that given the level of significance is 0.1.

We can describe our null and alternative hypotheses as:

- $H_0 : \mu_{non\ G8\ countries} \leq 1250$
- $H_A : \mu_{non\ G8\ countries} > 1250$

# Hypothesis Testing

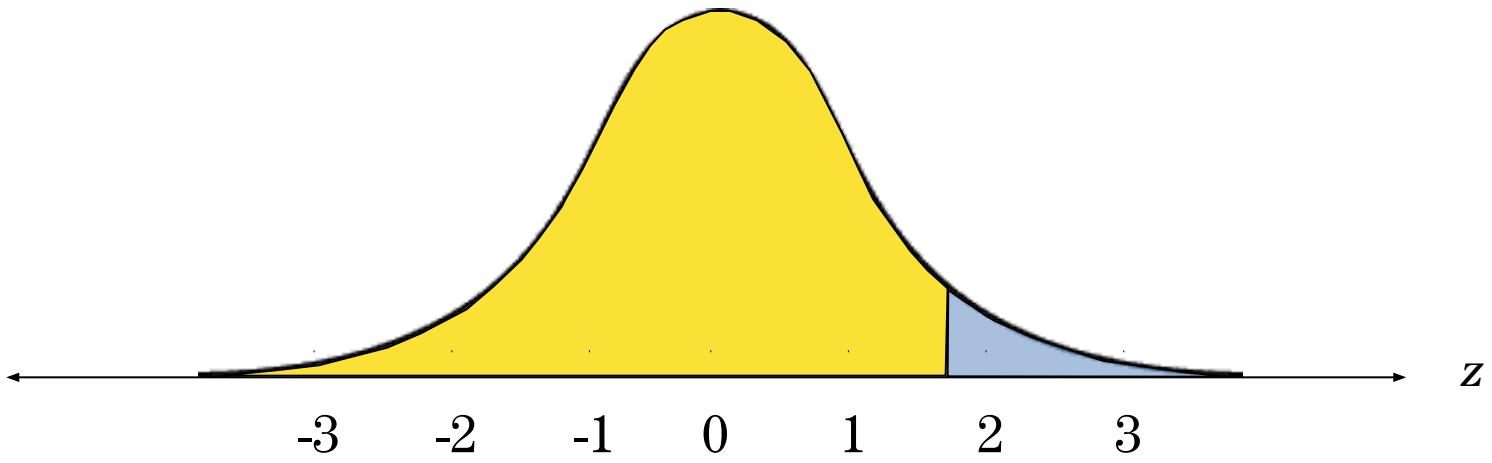
- $H_0 : \mu_{non\ G8\ countries} \leq 1250$
- $H_A : \mu_{non\ G8\ countries} > 1250$

Sample Statistic:

$$\bar{x} = 1400$$

$\mu$ = 1250, $\sigma$ = 700, n=49

Test Statistic: $z_{value} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{1400-1250}{700/\sqrt{49}} = 1.5$



# Hypothesis Testing

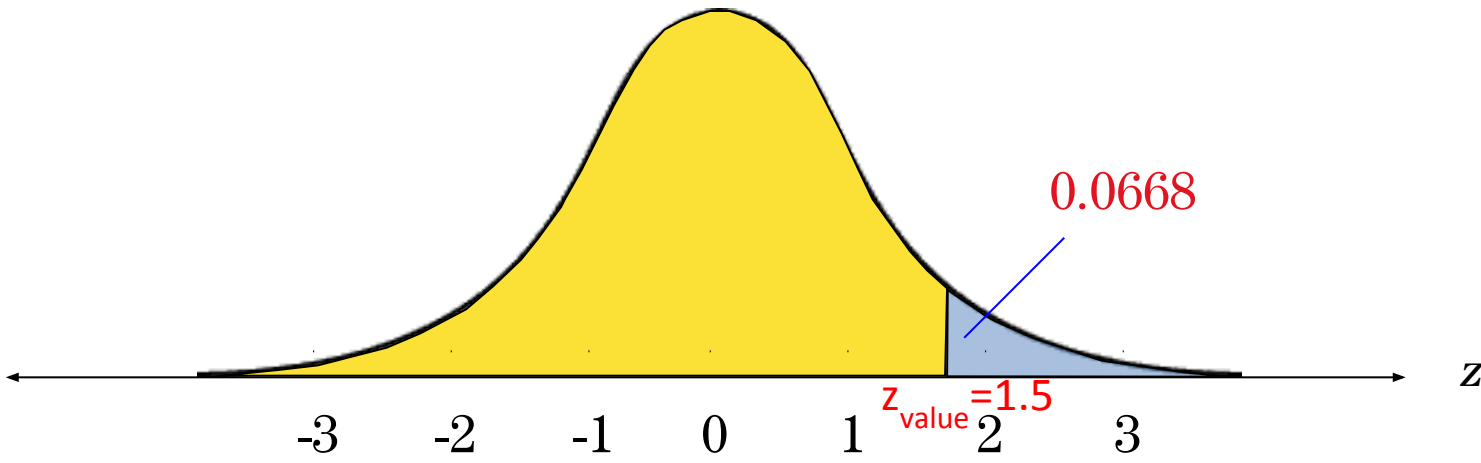
- $H_0 : \mu_{non\ G8\ countries} \leq 1250$
- $H_A : \mu_{non\ G8\ countries} > 1250$

Sample Statistic: $\bar{x} = 1400$

$\mu$ = 1250, $\sigma$ = 700

*Test Statistic:* $z_{value} = \frac{\bar{x} - \mu}{\sigma} = \frac{1400 - 1250}{700} = 1.5$

$$p\ value = P(\bar{Z} > 1.5) = 1 - 0.9332 = 0.0668$$



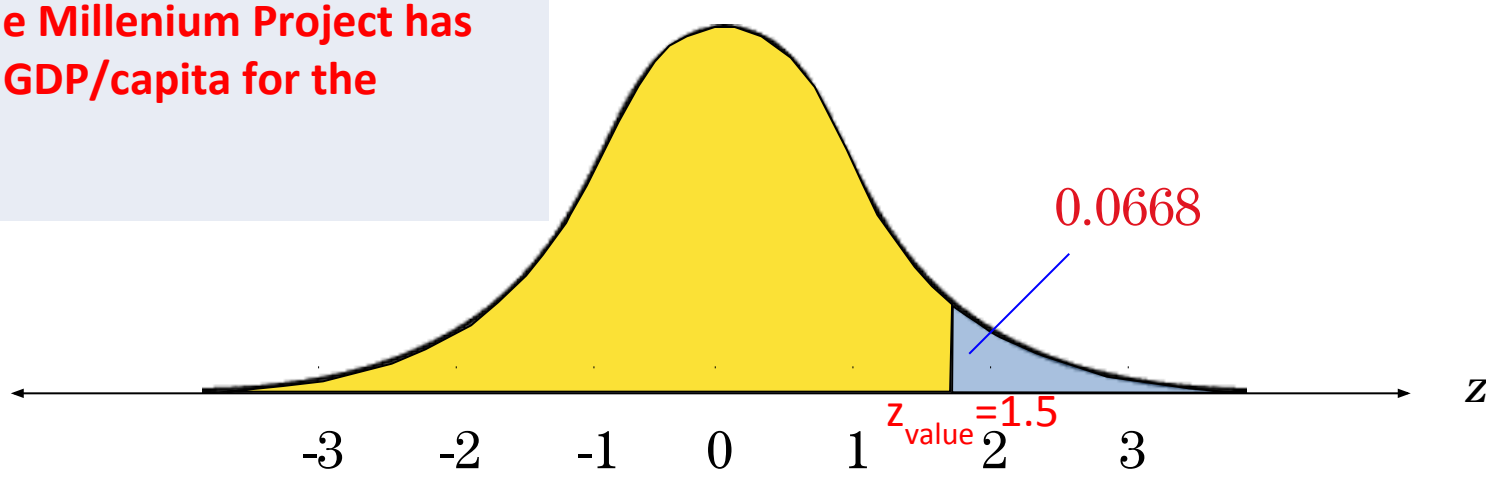
# Hypothesis Testing

Let's consider an example from the United Nations' Millennium Project. Countries outside of the G8 (the group of the eight most economically powerful nations) typically have an average per capita gross domestic product (GDP) of $1,250 or less. Now the UN has claimed that its Millennium Project has substantially improved the average GDP/Capita for these countries. In fact, a random sample of 49 non-G8 countries revealed an average GDP/capita of $1,400 with a standard deviation of $700. Did the Millennium Project increase average GDP/capita for non-G8 countries? In this case, let's arbitrarily allow for only a 10% probability that we will randomly sample and get a mean GDP/capita that is significantly higher than $1,250.

- $H_0 : \mu_{non\ G8\ countries} \leq 1250$
- $H_A : \mu_{non\ G8\ countries} > 1250$

$$p\ value = P(\bar{Z} > 1.5) = 1 - 0.9332 = 0.0668$$

**Since p value = 0.0668 < 0.10, we reject the null hypothesis, and agree with the UN on their claim that the Millenium Project has increased the avg GDP/capita for the non-G8 countries.**



# Module 5: Bivariate Analysis - Regression

# Topics covered:

- What is Bivariate analysis
- Correlation
  - Correlation coefficient
- Linear regression
  - Finding the regression line
  - Standard error of estimate

# Bivariate analysis

- Statistics of the relationship between two variables, including how to summarize the relationship, graph it, make estimations, build interval estimates using the relationship
- Restrict our analyses to continuous numerical variables (not categorical variables such as gender, race, ethnicity, etc.) and
- Focus exclusively on linear relationships between variables

## Bivariate Data

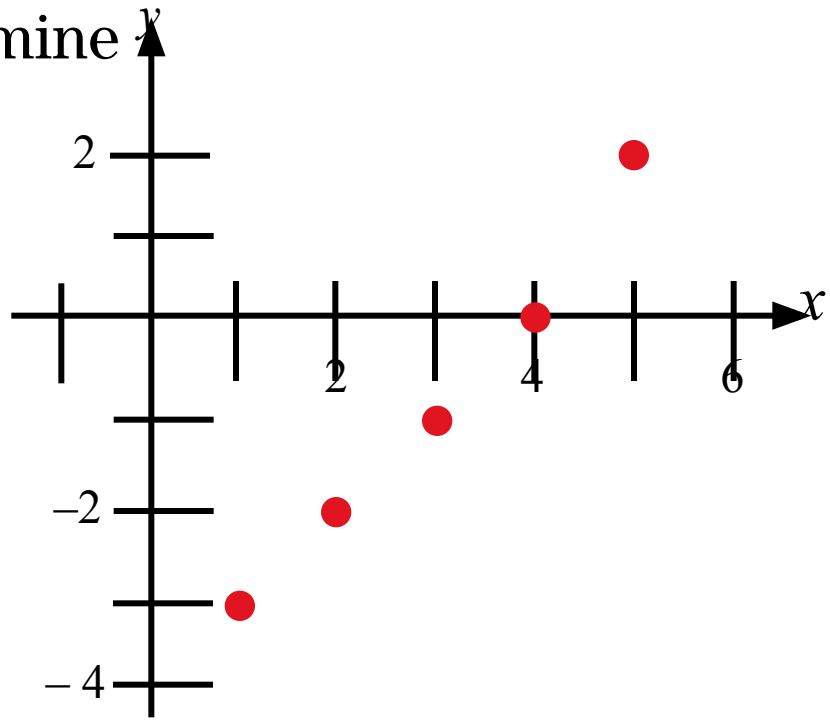
- paired data (x,y)
- for example, the heights and weights of different people, or their salaries and years of education.

**Rule #1 of statistics: GRAPH THE DATA!**

# Correlation

A **correlation** is a relationship between two variables. The data can be represented by the ordered pairs (x, y) where x is the **independent** (or **explanatory**) **variable**, and y is the **dependent** (or **response**) **variable**.

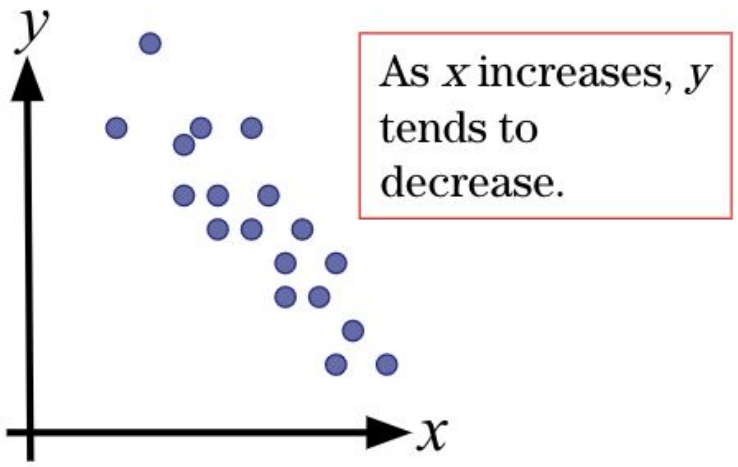
A **scatter plot** can be used to determine whether a linear (straight line) correlation exists between two variables.



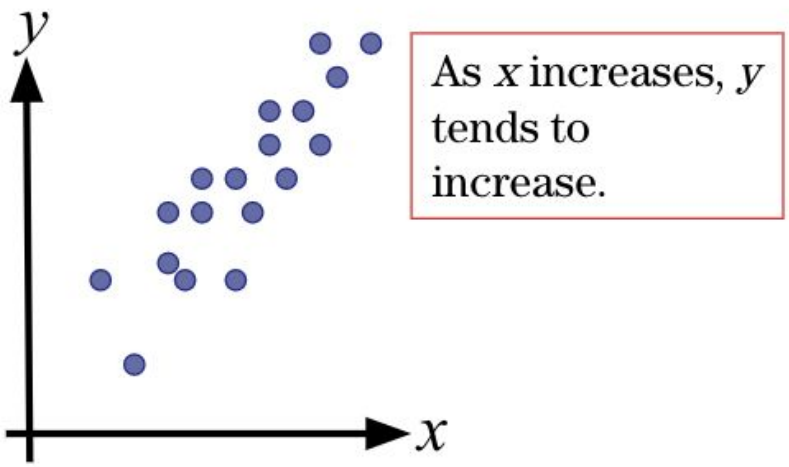

**Example**:

| $x$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $y$ | -3 | -2 | -1 | 0 | 2 |

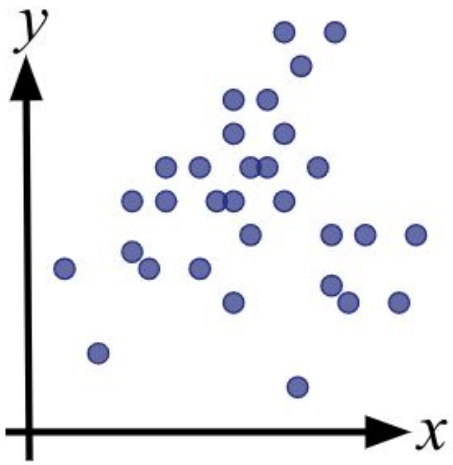
# Correlation



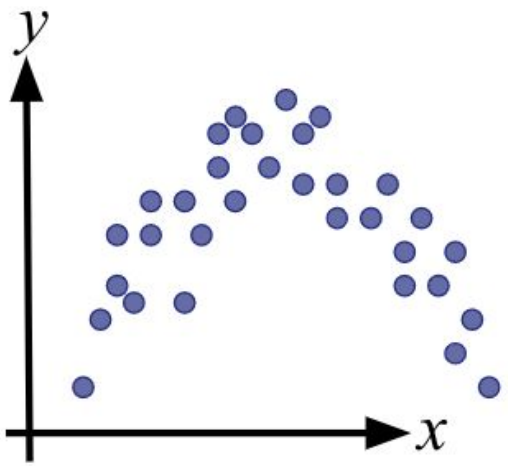

Negative Linear Correlation




Positive Linear Correlation

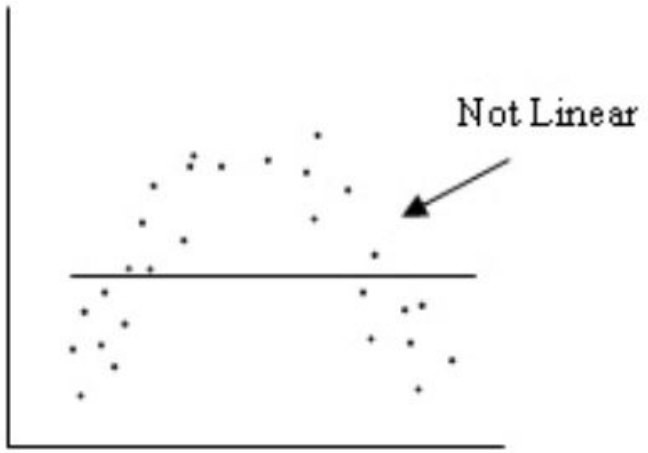


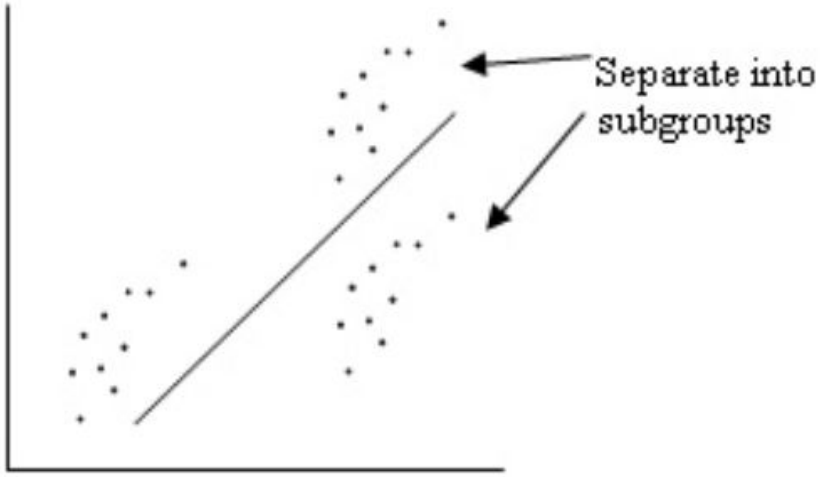

No Correlation

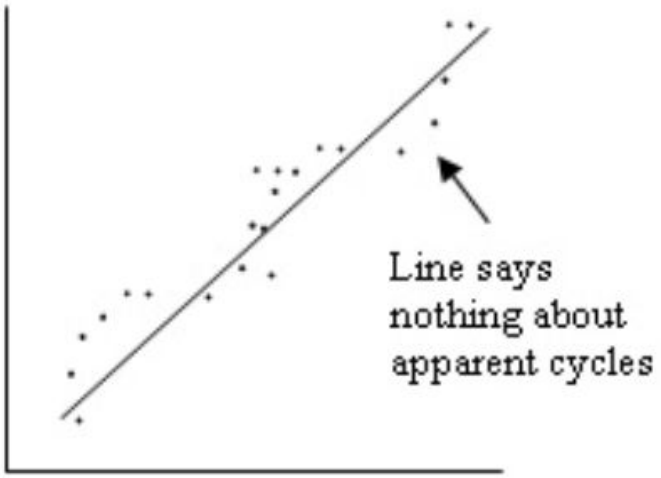


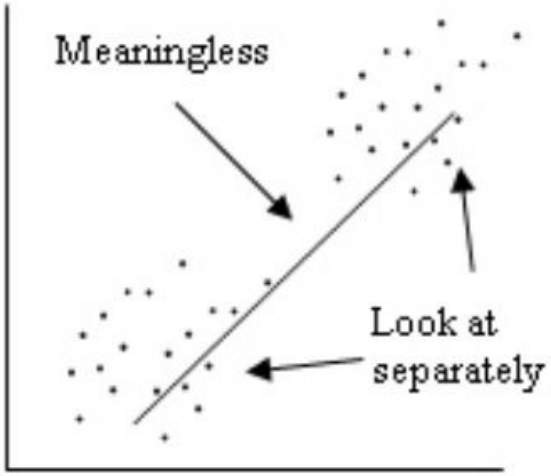

Nonlinear Correlation

# Correlation



# Correlation Coefficient

The **correlation coefficient** is a measure of the strength and the direction of a linear relationship between two variables. The symbol *r* represents the sample correlation coefficient. The formula for *r* is

$$r = \frac{1}{n-1} \sum \frac{(x-\bar{x})(y-\bar{y})}{s_x s_y}$$

- The range of the correlation coefficient, r is –1 to 1.
- If *x* and *y* have a strong positive linear correlation, *r* is close to 1.
- If *x* and *y* have a strong negative linear correlation, *r* is close to –1.
- If there is no linear correlation or a weak linear correlation, *r* is close to 0.

# Correlation Coefficient

**Example**:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Display the scatter plot.

b.) Calculate the correlation coefficient *r*.

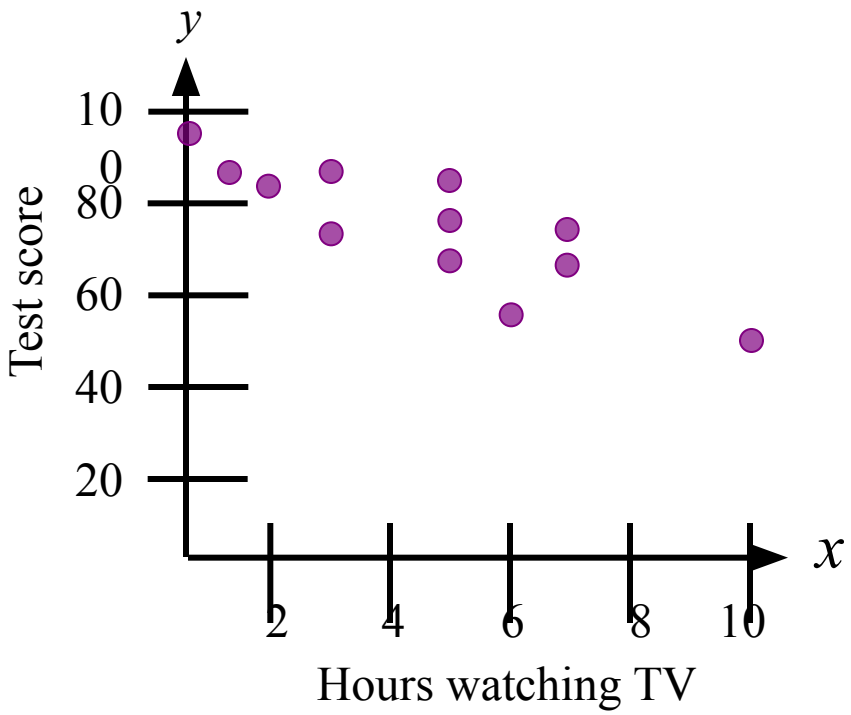
| Hours, $x$ | 0 | 1 | 2 | 3 | 3 | 5 | 5 | 5 | 6 | 7 | 7 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test score, $y$ | 96 | 85 | 82 | 74 | 95 | 68 | 76 | 84 | 58 | 65 | 75 | 50 |

Continued.

# Correlation Coefficient

**Example continued**:

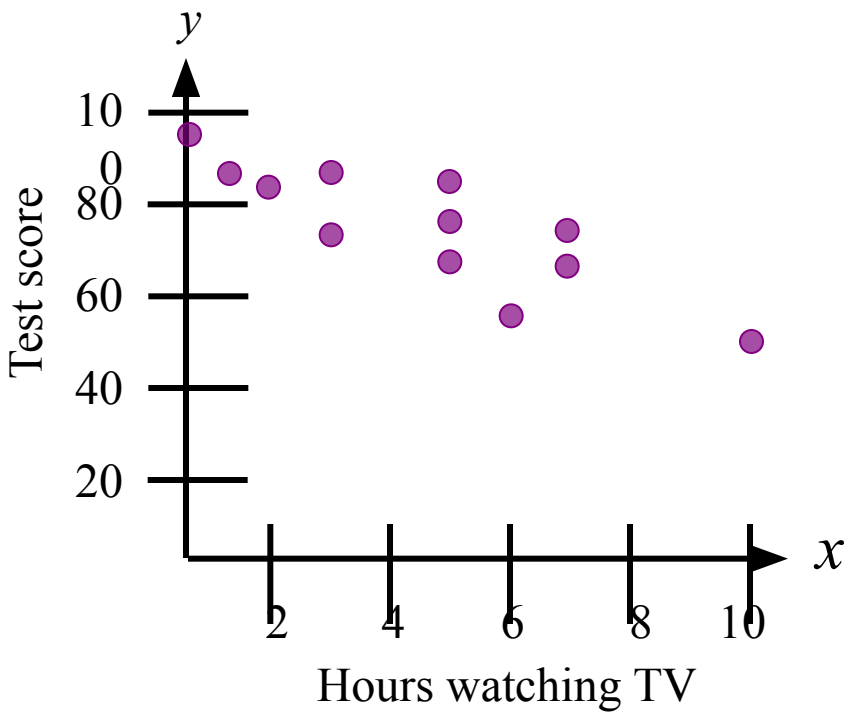
| Hours, $x$ | 0 | 1 | 2 | 3 | 3 | 5 | 5 | 5 | 6 | 7 | 7 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test score, $y$ | 96 | 85 | 82 | 74 | 95 | 68 | 76 | 84 | 58 | 65 | 75 | 50 |




**r = -0.831**

# Linear Regression

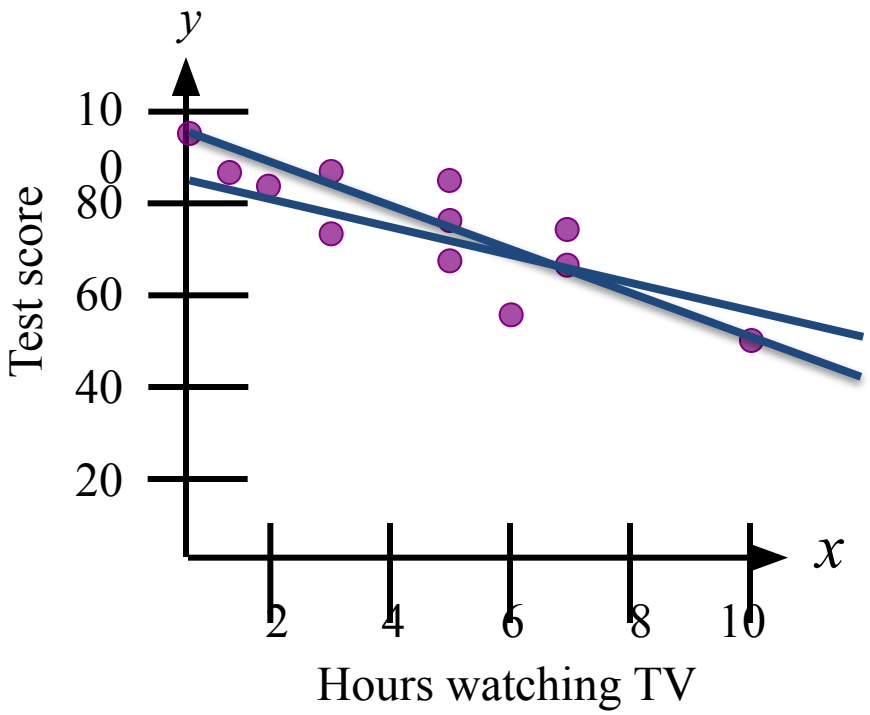
- approximates the relationship between two variables using a straight line.
- a scatterplot helps decide if a linear model is suitable




**r = -0.831**

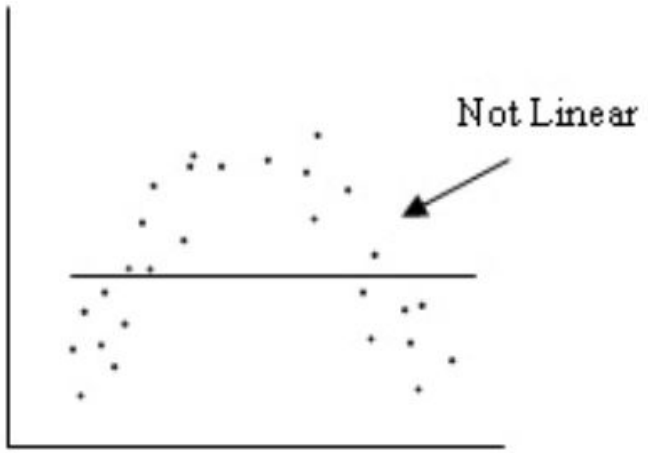
# Linear Regression

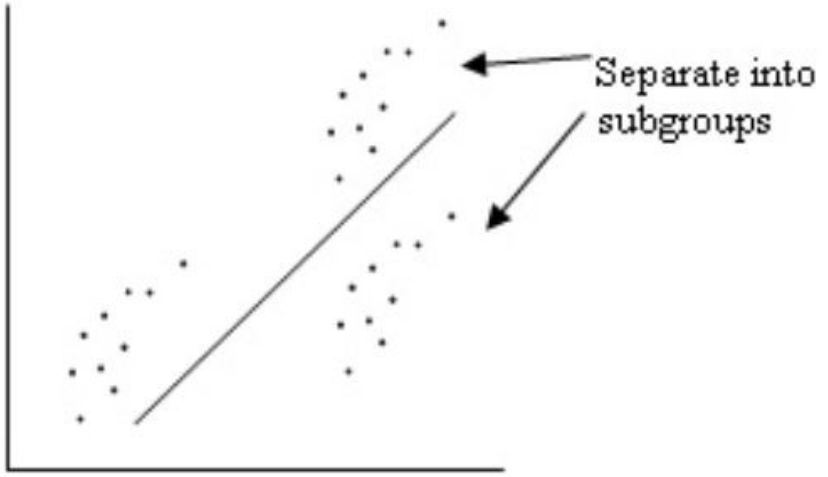
- approximates the relationship between two variables using a straight line.
- a scatterplot helps decide if a linear model is suitable

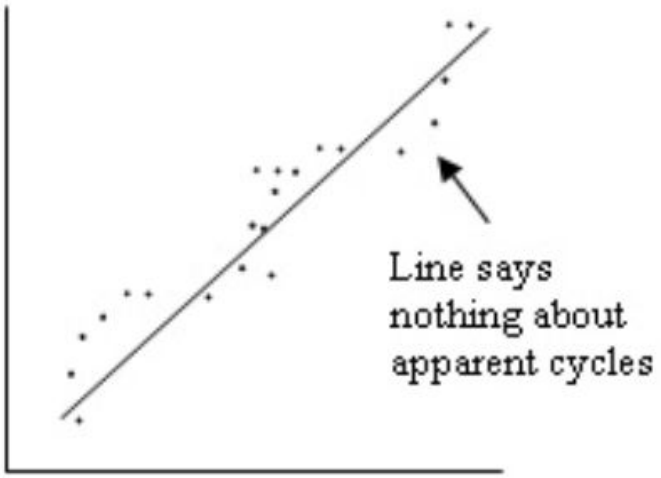


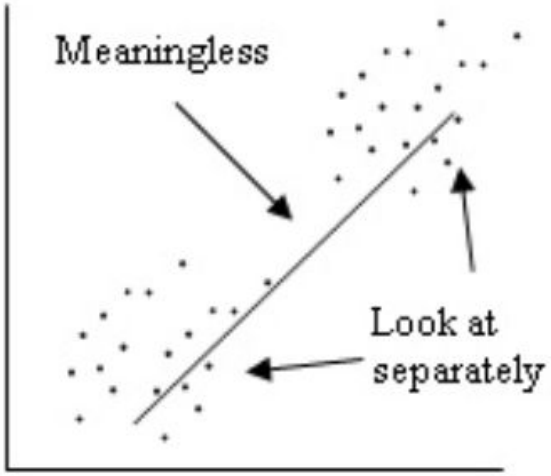

**r = -0.831**

# Correlation



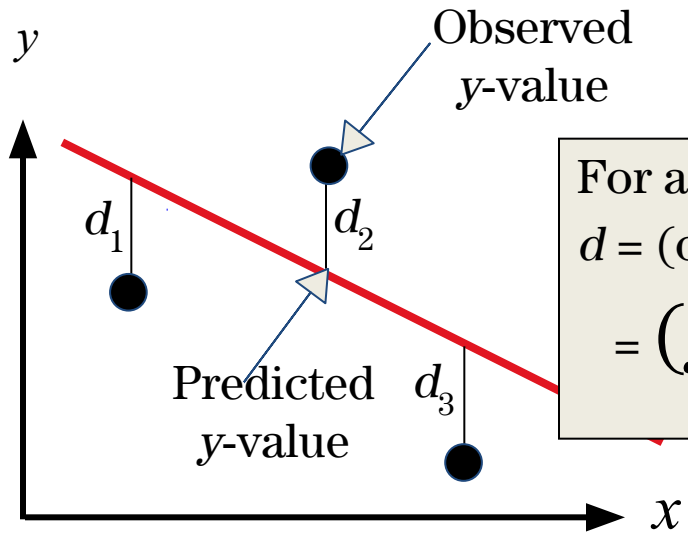



# Residuals

After verifying that the linear correlation between two variables is significant, next we determine the equation of the line that can be used to predict the value of *y* for a given value of *x*.




For a given *x*-value,

*d* = (observed *y*-value) – (predicted *y*-value)

$= (y - \hat{y})$

Each data point $d_i$ represents the difference between the observed *y*-value and the predicted *y*-value. $d_i$ is called the **residuals**.

# Finding the Regression Line

A **regression line** is the line for which the sum of the squares of the residuals is a minimized.

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (\text{observed y} - \text{predicted y})^2$$

**The Equation of a Regression Line**

The equation of a regression line for an independent variable $x$ and a dependent variable $y$ is

$$\hat{y} = mx + b$$

where $\hat{y}$ is the predicted $y$-value for a given $x$-value. The slope $m$ and $y$-intercept $b$ are given by

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \text{ and } b = \bar{y} - m\bar{x} = \frac{\sum y}{n} - m\frac{\sum x}{n}$$

where $\bar{y}$ is the mean of the y-values and $\bar{x}$ is the mean of the $x$-values. The regression line always passes through $(\bar{x}, \bar{y})$.

# Regression Line

**Example**:

Find the equation of the regression line.

| $x$ | $y$ | $xy$ | $x^2$ | $y^2$ |
|---|---|---|---|---|
| 1 | $-3$ | $-3$ | 1 | 9 |
| 2 | $-1$ | $-2$ | 4 | 1 |
| 3 | 0 | 0 | 9 | 0 |
| 4 | 1 | 4 | 16 | 1 |
| 5 | 2 | 10 | 25 | 4 |
| $\sum x = 15$ | $\sum y = -1$ | $\sum xy = 9$ | $\sum x^2 = 55$ | $\sum y^2 = 15$ |

$$m = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(9) - (15)(-1)}{5(55) - (15)^2} = \frac{60}{50} = 1.2$$

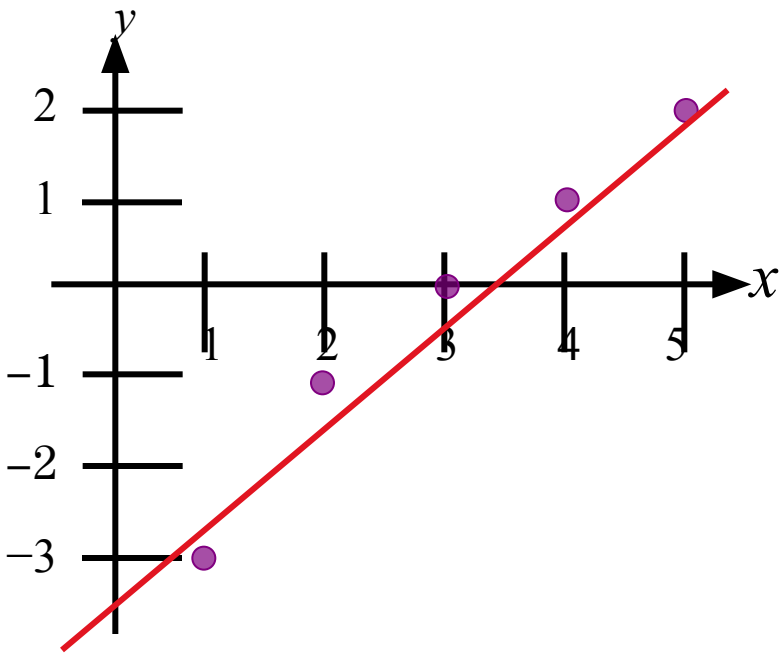
# Regression Line

**Example continued**:

$$b = y - mx = \frac{-1}{5} - (1.2)\frac{15}{5} = -3.8$$

The equation of the regression line is

$$y = 1.2x - 3.8.$$



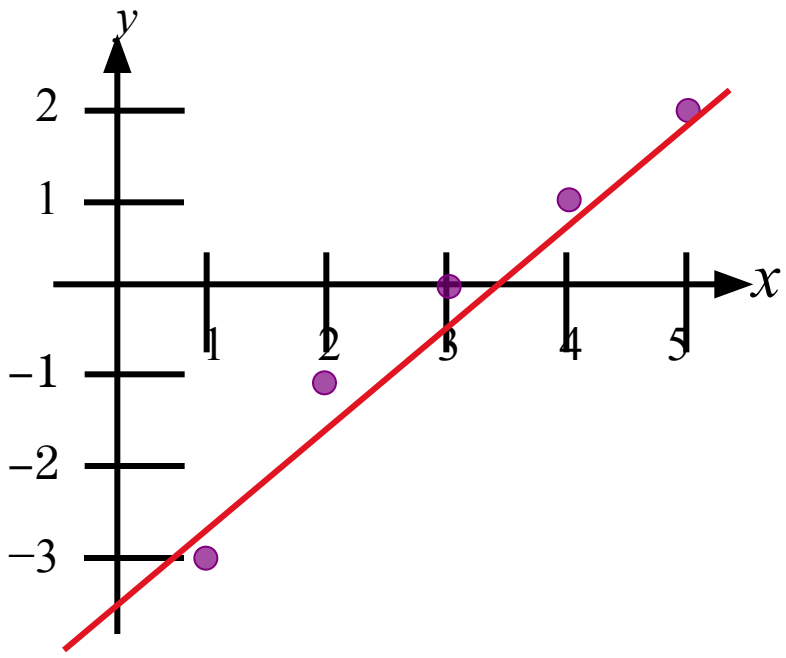
# Regression Line

**Predict the y for x=1.5**

The equation of the regression line is

$$y = 1.2x - 3.8.$$

Plug **x=1.5** to the above equation and find y



# Regression Line

Some things to remember!

- Avoid Extrapolation! Do not use linear regression to predict values of y for values of x outside the range of the data.
- Correlation does not imply causation! If a scatterplot shows a definite pattern, and the data are found to have a strong correlation, that doesn't necessarily mean that a cause-and-effect relationship exists between the two variables.

# Regression Line

**Example**:
The following data represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday.

a.) Find the equation of the regression line.

b.) Use the equation to find the expected test score for a student who watches 9 hours of TV.

| Hours, $x$ | 0 | 1 | 2 | 3 | 3 | 5 | 5 | 5 | 6 | 7 | 7 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test score, $y$ | 96 | 85 | 82 | 74 | 95 | 68 | 76 | 84 | 58 | 65 | 75 | 50 |
| $xy$ | 0 | 85 | 164 | 222 | 285 | 340 | 380 | 420 | 348 | 455 | 525 | 500 |
| $x^2$ | 0 | 1 | 4 | 9 | 9 | 25 | 25 | 25 | 36 | 49 | 49 | 100 |
| $y^2$ | 9216 | 7225 | 6724 | 5476 | 9025 | 4624 | 5776 | 7056 | 3364 | 4225 | 5625 | 2500 |

$\sum x = 54$ $\sum y = 908$ $\sum xy = 3724$ $\sum x^2 = 332$ $\sum y^2 = 70836$

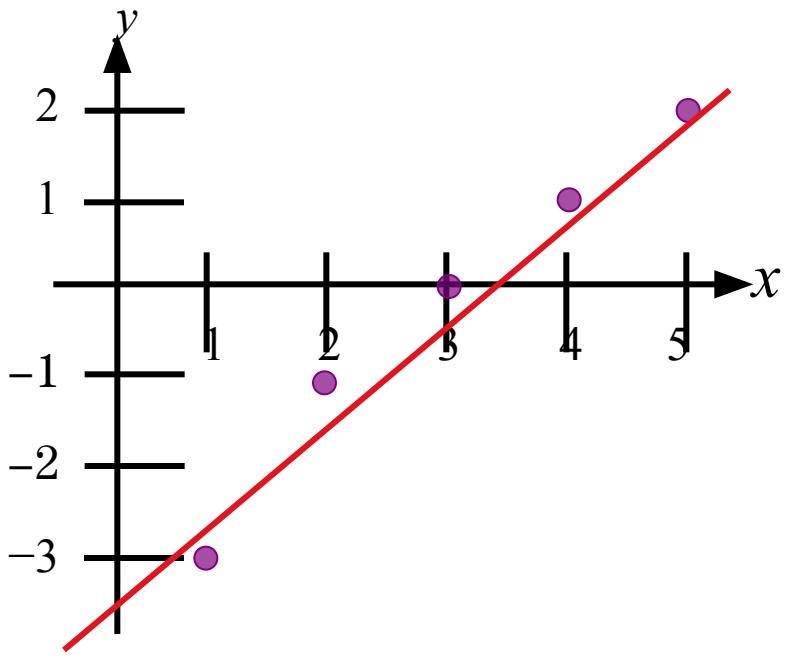
# Regression Line

**Predict the y for x=1.5**

The equation of the regression line is

$$y = 1.2x - 3.8.$$

Plug **x=1.5** to the above equation and find y



# The Standard Error of Estimate

The **standard error of estimate $s_e$** is the standard deviation of the observed $y_i$-values about the predicted $\hat{y}$-value for a given $x_i$-value. It is given by

$$s_e = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$$

where $n$ is the number of ordered pairs in the data set.

The closer the observed $y$-values are to the predicted $y$-values, the smaller the standard error of estimate will be.

# The Standard Error of Estimate

**Example**:
The regression equation for the data that represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday is

$\hat{y} = -4.07x + 93.97.$

Find the standard error of estimate.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Hours, $x_i$ | | 0 | 1 | 2 | 3 | 3 | 5 |
| Test score, $y_i$ | | 96 | 85 | 82 | 74 | 95 | 68 |
| $\hat{y}_i$ | | 93.97 | 89.9 | 85.83 | 81.76 | 81.76 | 73.62 |
| $(y_i - \hat{y}_i)^2$ | | 4.12 | 24.01 | 14.67 | 60.22 | 175.3 | 31.58 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Hours, $x_i$ | 5 | 5 | 6 | 7 | 7 | 10 |
| Test score, $y_i$ | 76 | 84 | 58 | 65 | 75 | 50 |
| $\hat{y}_i$ | 73.62 | 73.62 | 69.55 | 65.48 | 65.48 | 53.27 |
| $(y_i - \hat{y}_i)^2$ | 5.66 | 107.74 | 133.4 | 0.23 | 90.63 | 10.69 |

Continued.

# The Standard Error of Estimate

**Example**:

The regression equation for the data that represents the number of hours 12 different students watched television during the weekend and the scores of each student who took a test the following Monday is

$\hat{y} = -4.07x + 93.97.$

Find the standard error of estimate.

| Hours, $x_i$ | 0 | 1 | 2 | 3 | 3 | 5 |
|---|---|---|---|---|---|---|
| Test score, $y_i$ | 96 | 85 | 82 | 74 | 95 | 68 |
| $\hat{y}_i$ | 93.97 | 89.9 | 85.83 | 81.76 | 81.76 | 73.62 |
| $(y_i - \hat{y}_i)^2$ | 4.12 | 24.01 | 14.67 | 60.22 | 175.3 | 31.58 |

| Hours, $x_i$ | 5 | 5 | 6 | 7 | 7 | 10 |
|---|---|---|---|---|---|---|
| Test score, $y_i$ | 76 | 84 | 58 | 65 | 75 | 50 |
| $\hat{y}_i$ | 73.62 | 73.62 | 69.55 | 65.48 | 65.48 | 53.27 |
| $(y_i - \hat{y}_i)^2$ | 5.66 | 107.74 | 133.4 | 0.23 | 90.63 | 10.69 |

Continued.

# The Standard Error of Estimate

**Example continued**:

$$\sum(y_i - \hat{y}_i)^2 = 658.25$$

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{658.25}{12-2}} \approx 8.11$$

The standard deviation of the student test scores for a specific number of hours of TV watched is about 8.11.

# References


- Ramsey, D., Statistics Essentials for Dummies.
- Thurman, P., Hamilton, A., Hammer, C., Data Analysis and Decision-Making in Fragile Contexts.