

# Unit 3–Inferential Statistics

BY: Padmavati Sarode

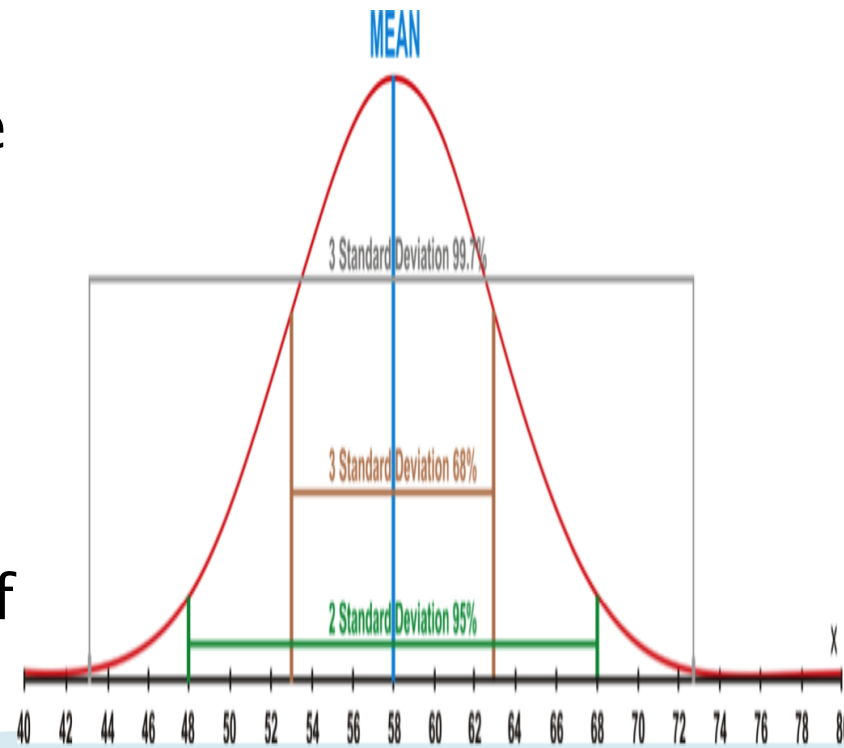
# Distribution in Inferential Statistics

inferential statistics: A branch of mathematics that involves drawing conclusions about a population based on sample data drawn from it.

sampling distribution: The probability distribution of a given statistic based on a random sample.

# Normal Distribution in inferential statistics

The **normal distribution** is a symmetrical, bell-shaped **distribution** in which the mean, median and mode are all equal. It is a central component of **inferential statistics**. The standard **normal distribution** is a **normal distribution** represented in z scores. It always has a mean of zero and a standard deviation of one.



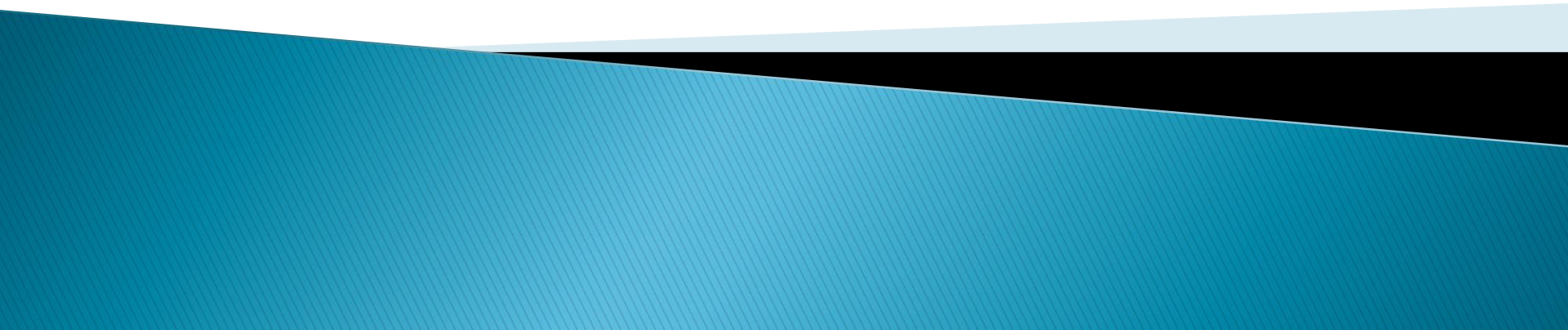
# What is a normal distribution in statistics?

- ▶ The **normal distribution** is a continuous probability **distribution** that is symmetrical on both sides of the mean, so the right side of the center is a mirror image of the left side.  
... The **normal distribution** is often called the bell curve because the graph of its probability density looks like a bell.

# How do you determine normal distribution?

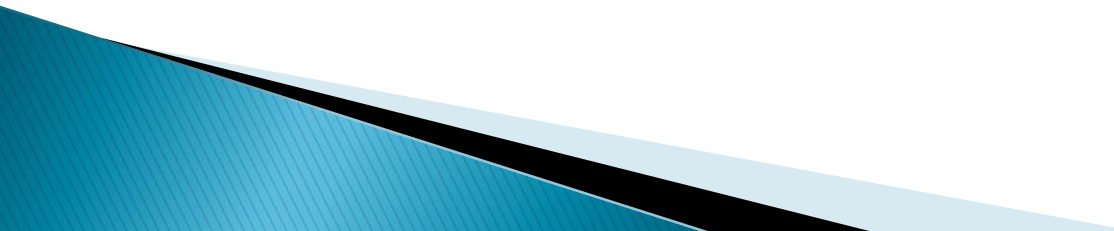
Explanation: A **normal distribution** is one in which the values are evenly distributed both above and below the mean.

A population has a precisely **normal distribution** if the mean, mode, and median are all equal. For the population of 3,4,5,5,5,6,7, the mean, mode, and median are all 5.





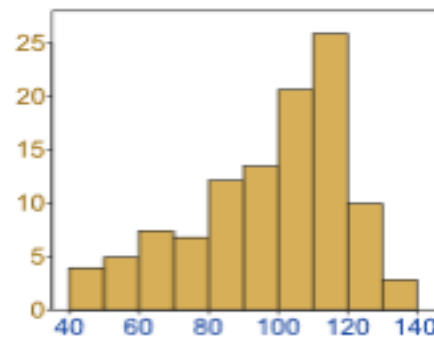
# What is a standard normal distribution in statistics?

- ▶ The **standard normal distribution** is a special case of the **normal distribution** . It is the **distribution** that occurs when a **normal** random variable has a mean of zero and a **standard** deviation of one. The **normal** random variable of a **standard normal distribution** is called a **standard** score or a z score.
- 

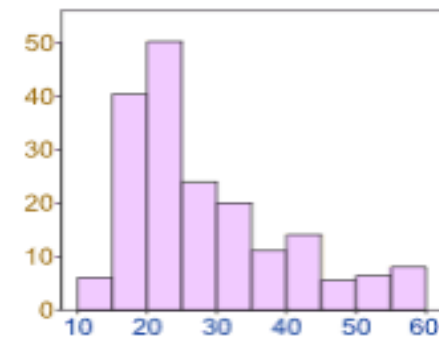
# Normal Distribution

Data can be "distributed" (spread out) in different ways.

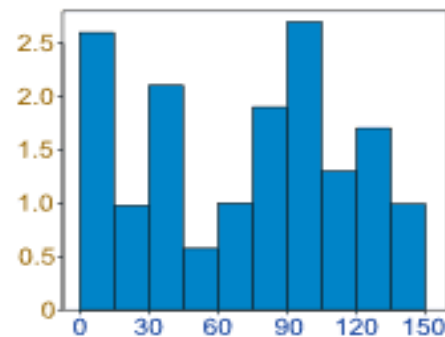
It can be spread out  
more on the left



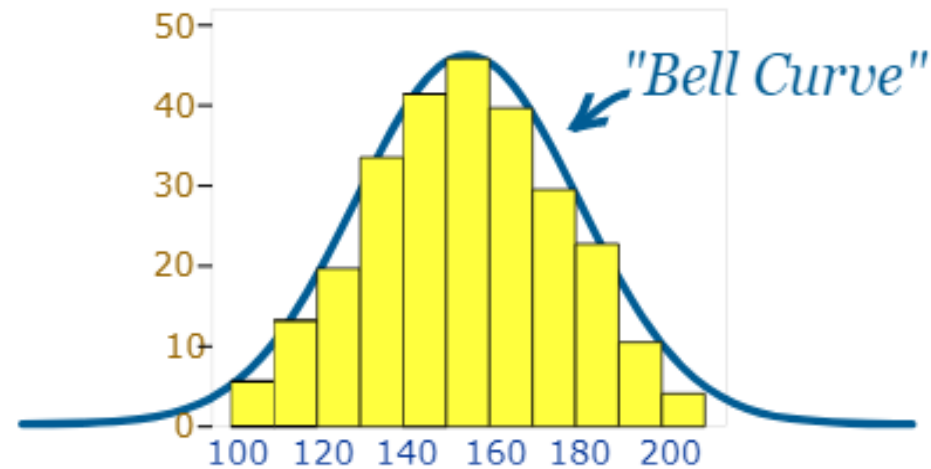
Or more on the right



Or it can be all jumbled up



But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



A Normal Distribution

The "Bell Curve" is a Normal Distribution.  
And the yellow histogram shows some data that follows it closely, but not perfectly (which is usual).



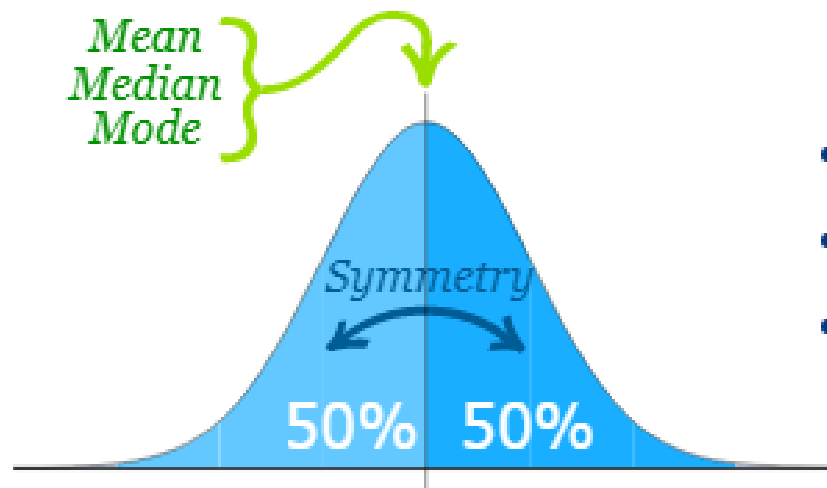
It is often called a "Bell Curve" because it looks like a bell.



Many things closely follow a Normal Distribution:

- heights of people
- size of things produced by machines
- errors in measurements
- blood pressure
- marks on a test

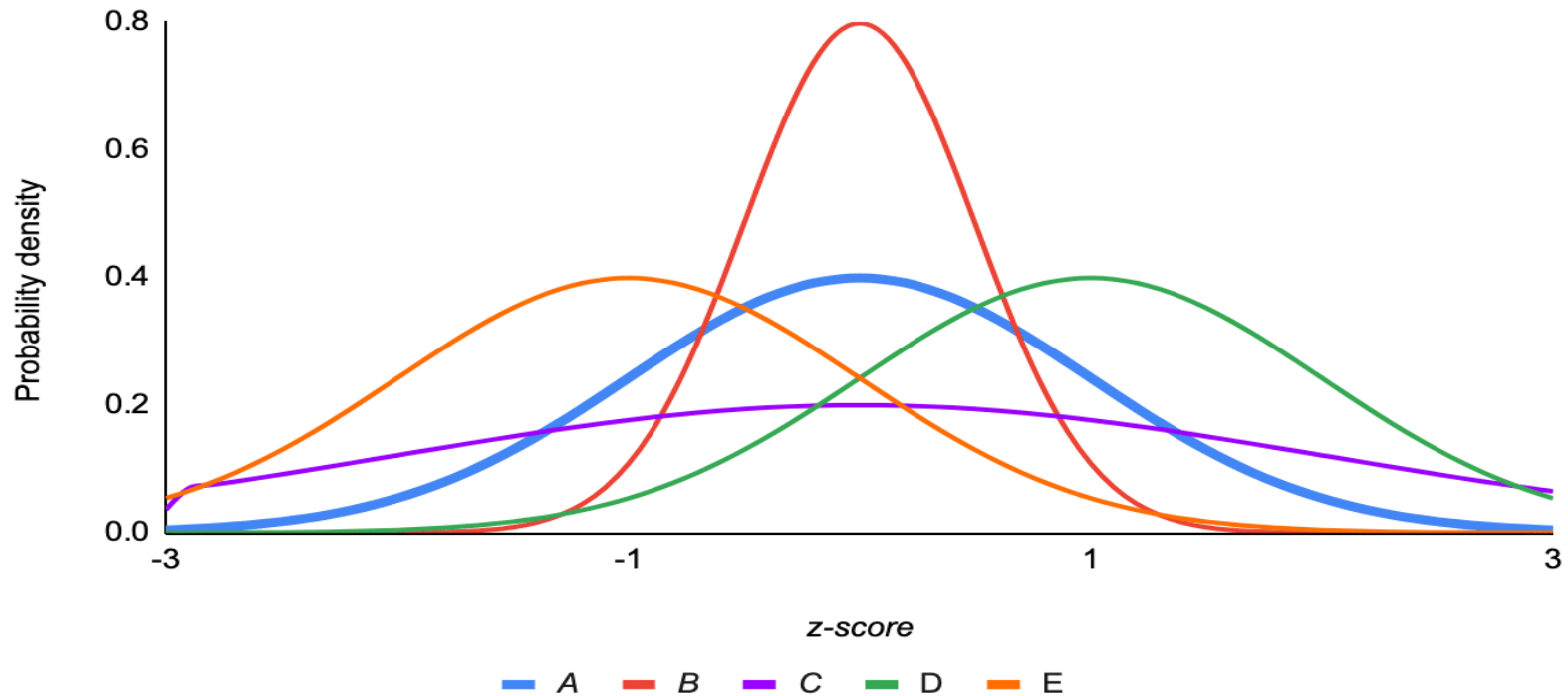
We say the data is "normally distributed":



The **Normal Distribution** has:

- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

## Normal distributions



Curve

A ( $M = 0$ ,  $SD = 1$ )

B ( $M = 0$ ,  $SD = 0.5$ )

C ( $M = 0$ ,  $SD = 2$ )

D ( $M = 1$ ,  $SD = 1$ )

E ( $M = -1$ ,  $SD = 1$ )

Position or shape (relative to standard normal distribution)

Standard normal distribution

Squeezed, because  $SD < 1$

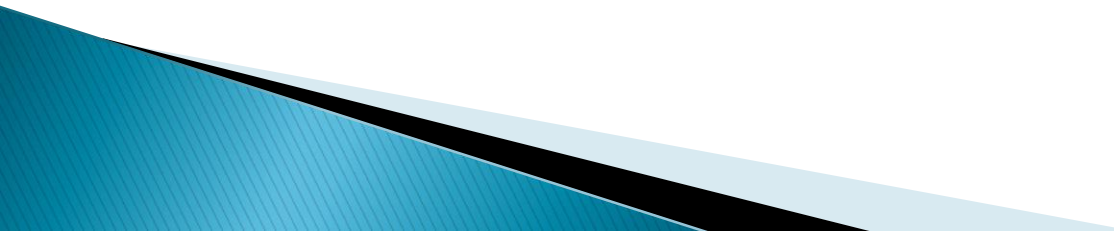
Stretched, because  $SD > 1$

Shifted right, because  $M > 0$

Shifted left, because  $M < 0$

# What is the difference between normal distribution and standard normal distribution?

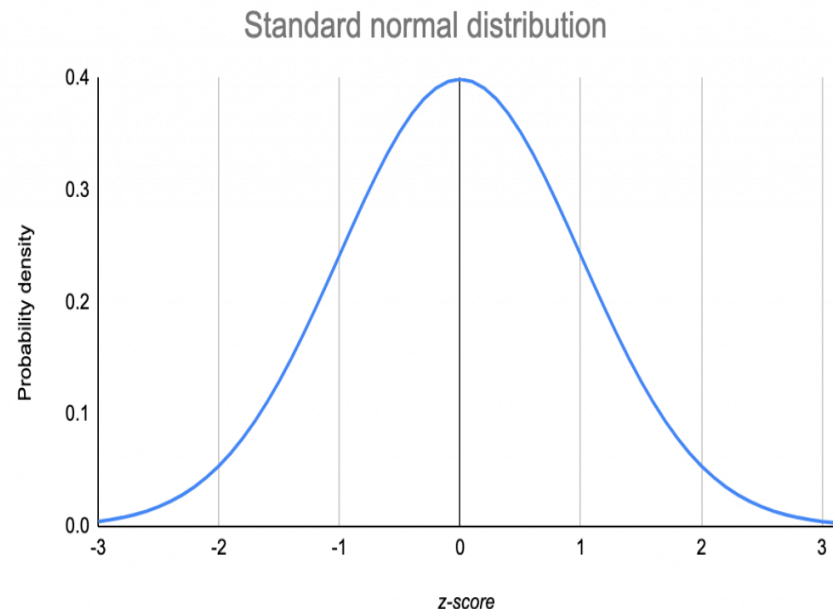
A **normal distribution** is determined by two parameters the mean and the variance. ... Now the **standard normal distribution** is a specific **distribution** with mean 0 and variance 1. This is the **distribution** that is used to construct tables of the **normal distribution**.



# The standard normal distribution

- The **standard normal distribution**, also called the **z-distribution**, is a special **normal distribution** where the **mean** is 0 and the **standard deviation** is 1.
- Any normal distribution can be standardized by converting its values into z-scores. Z-scores tell you how many standard deviations from the mean each value lies.

Converting a normal distribution into a z-distribution allows you to calculate the probability of certain values occurring and to compare different data sets.

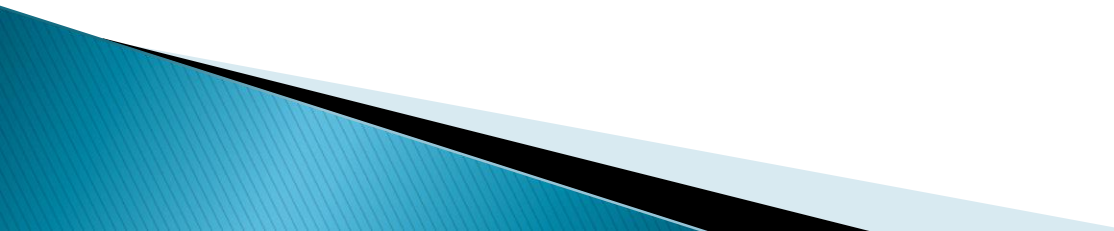


# Normal distribution vs the standard normal distribution

- All normal distributions, like the standard normal distribution, are unimodal and symmetrically distributed with a bell-shaped curve. However, a normal distribution can take on any value as its mean and standard deviation. In the standard normal distribution, the mean and standard deviation are always fixed.
- Every normal distribution is a version of the standard normal distribution that's been stretched or squeezed and moved horizontally right or left.
- The mean determines where the curve is centered. Increasing the mean moves the curve right, while decreasing it moves the curve left.
- The standard deviation stretches or squeezes the curve. A small standard deviation results in a narrow curve, while a large standard deviation leads to a wide curve.

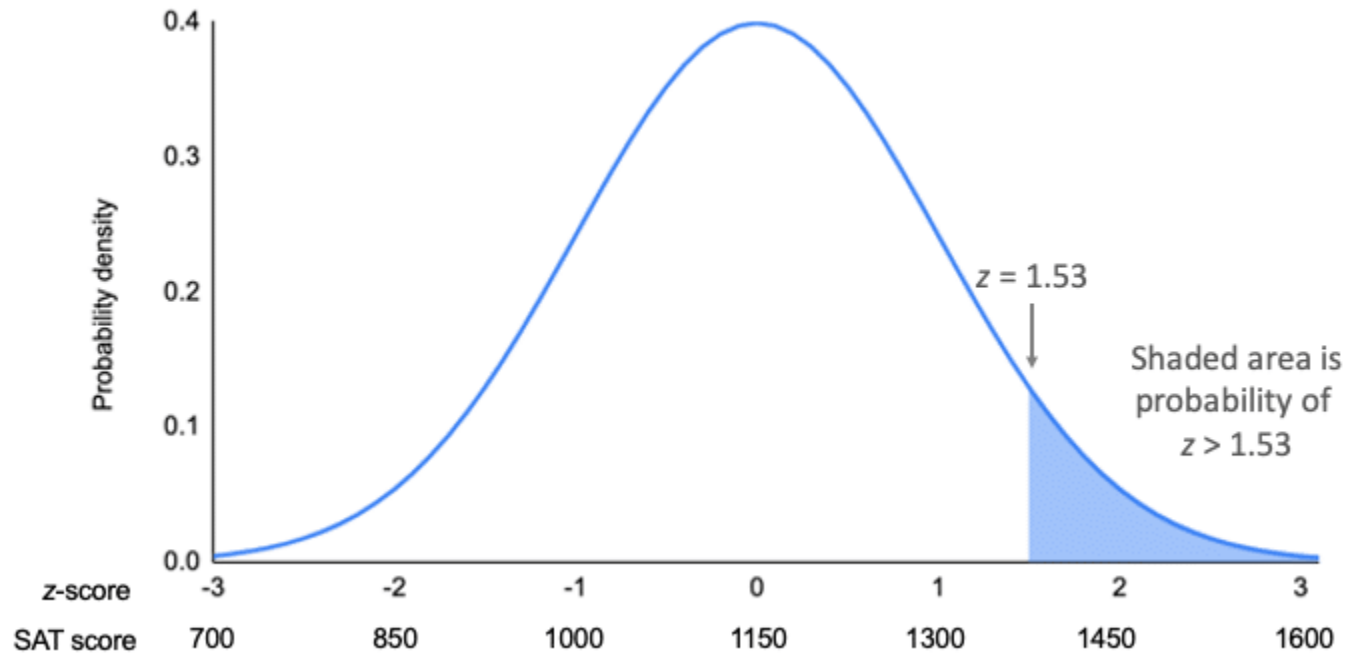
# What is the difference between normal distribution and standard normal distribution?

A **normal distribution** is determined by two parameters the mean and the variance. ... Now the **standard normal distribution** is a specific **distribution** with mean 0 and variance 1. This is the **distribution** that is used to construct tables of the **normal distribution**.



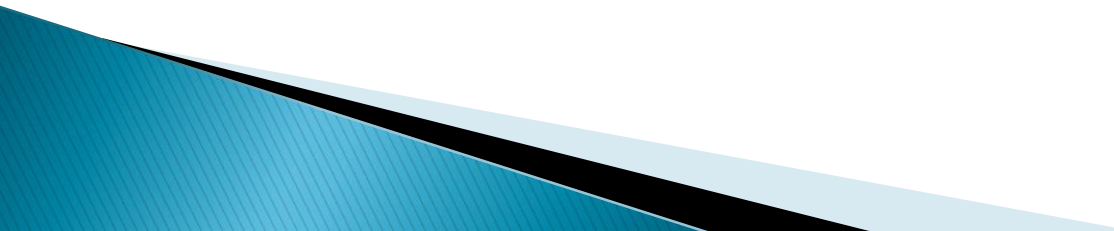


## Standard normal distribution



# What are the properties of the standard normal distribution?

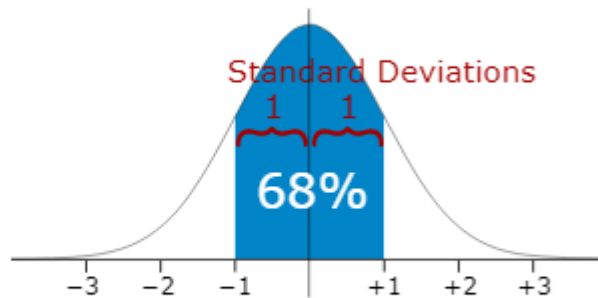
## Properties of a normal distribution

1. The mean, mode and median are all equal.
  2. The curve is symmetric at the center (i.e. around the mean,  $\mu$ ).
  3. Exactly half of the values are to the left of center and exactly half the values are to the right.
  4. The total **area** under the curve is 1.
- 

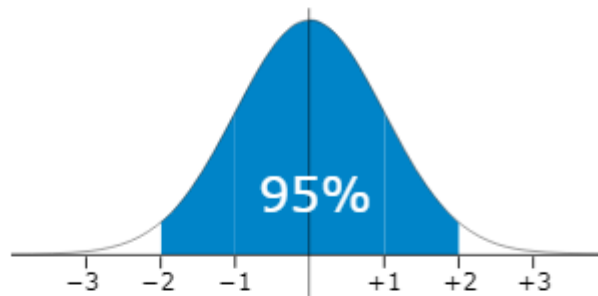
# The Standard Normal Distribution

The standard normal distribution is a normal distribution with a mean of zero and standard deviation of 1. The standard normal distribution is centered at zero and the degree to which a given measurement deviates from the mean is given by the standard deviation. For the standard normal distribution, 68% of the observations lie within 1 standard deviation of the mean; 95% lie within two standard deviation of the mean; and 99.9% lie within 3 standard deviations of the mean. To this point, we have been using "X" to denote the variable of interest (e.g.,  $X = \text{BMI}$ ,  $X = \text{height}$ ,  $X = \text{weight}$ ). However, when using a standard normal distribution, we will use "Z" to refer to a variable in the context of a standard normal distribution. After standardization, the  $\text{BMI} = 30$  discussed on the previous page is shown below lying 0.16667 units above the mean of 0 on the standard normal distribution on the right.

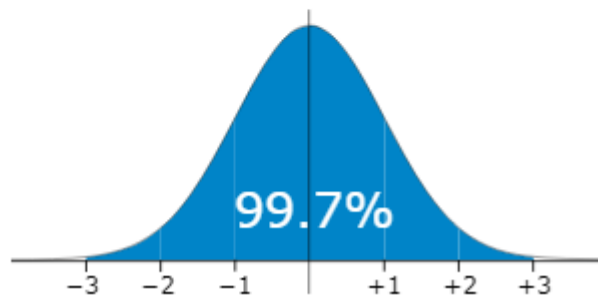
When we [calculate the standard deviation](#) we find that **generally**:



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



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



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

## What are the formulas for the standard deviation?

The **sample standard deviation formula** is:

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

where,

s = sample standard deviation

$\sum$  = sum of...

$\bar{X}$  = sample mean

n = number of scores in sample.

The **population standard deviation formula** is:

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{n}}$$

where,

$\sigma$  = population standard deviation

$\sum$  = sum of...

$\mu$  = population mean

n = number of scores in sample.

# Calculating standard deviation step by step

## Overview of how to calculate standard deviation

The formula for standard deviation (SD) is

$$SD = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

where  $\sum$  means "sum of",  $x$  is a value in the data set,  $\mu$  is the mean of the data set, and  $N$  is the number of data points in the population.

**Step 1:** Find the mean.

**Step 2:** For each data point, find the square of its distance to the mean.

**Step 3:** Sum the values from Step 2.

**Step 4:** Divide by the number of data points.

**Step 5:** Take the square root.





**Step 1:** Find the mean  $\mu$ .

$$\mu = \frac{6 + 2 + 3 + 1}{4} = \frac{12}{4} = 3$$

**Step 2:** Find the square of the distance from each data point to the mean  $|x - \mu|^2$ .

$x$	$ x - \mu ^2$
6	$ 6 - 3 ^2 = 3^2 = 9$
2	$ 2 - 3 ^2 = 1^2 = 1$
3	$ 3 - 3 ^2 = 0^2 = 0$
1	$ 1 - 3 ^2 = 2^2 = 4$

**Steps 3, 4, and 5:**

Steps 3, 4, and 5:

$$SD = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

$$= \sqrt{\frac{9 + 1 + 0 + 4}{4}}$$

$$= \sqrt{\frac{14}{4}} \quad \text{Sum the squares of the distances (Step 3).}$$

$$= \sqrt{3.5} \quad \text{Divide by the number of data points (Step 4).}$$

$$\approx 1.87 \quad \text{Take the square root (Step 5).}$$

**Try it yourself**

And here's a data set:

1, 4, 7, 2, 6

Find the mean

$$\mu = \frac{1 + 4 + 7 + 2 + 6}{5} = \frac{20}{5} = 4$$

Find the square of the distances from each of the data points to the mean

$x$	$ x - \mu ^2$
1	$ 1 - 4 ^2 = 3^2 = 9$
4	$ 4 - 4 ^2 = 0^2 = 0$
7	$ 7 - 4 ^2 = 3^2 = 9$
2	$ 2 - 4 ^2 = 2^2 = 4$
6	$ 6 - 4 ^2 = 2^2 = 4$

Apply the formula

$$SD = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

$$= \sqrt{\frac{9 + 0 + 9 + 4 + 4}{5}}$$

$$= \sqrt{\frac{26}{5}}$$

$$= \sqrt{5.2}$$

**The answer**  
**The standard deviation is**  
**approximately 2.28**

**Example: 95% of students at school are between 1.1m and 1.7m tall.**

Assuming this data is **normally distributed** can you calculate the mean and standard deviation?

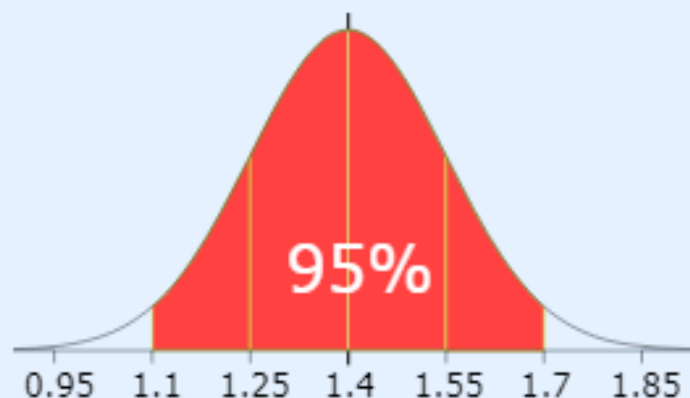
The mean is halfway between 1.1m and 1.7m:

$$\text{Mean} = (1.1\text{m} + 1.7\text{m}) / 2 = \mathbf{1.4\text{m}}$$

95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so:

$$\begin{aligned} 1 \text{ standard deviation} &= (1.7\text{m} - 1.1\text{m}) / 4 \\ &= 0.6\text{m} / 4 \\ &= \mathbf{0.15\text{m}} \end{aligned}$$

And this is the result:





## How to calculate a z-score

To standardize a value from a normal distribution, convert the individual value into a z-score:

1. Subtract the mean from your individual value.
2. Divide the difference by the standard deviation.

### Z-score formula

$$Z = \frac{x - \mu}{\sigma}$$

### Explanation

- $x$  = individual value
- $\mu$  = mean
- $\sigma$  = standard deviation

### Example: Finding a $z$ -score

You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score ( $M$ ) of 1150 and a standard deviation ( $SD$ ) of 150. You want to find the probability that SAT scores in your sample exceed 1380.

To standardize your data, you first find the  $z$ -score for 1380. The  $z$ -score tells you how many standard deviations away 1380 is from the mean.

Step 1: Subtract the mean from the  $x$  value.

$$x = 1380$$

$$M = 1150$$

$$x - M = 1380 - 1150 = 230$$

Step 2: Divide the difference by the standard deviation.

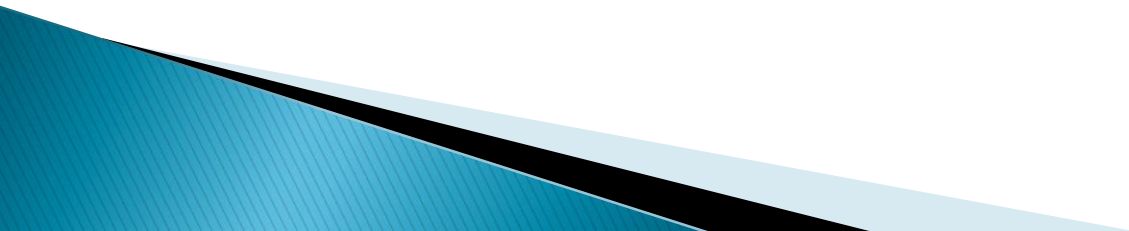
$$SD = 150$$

$$z = 230 \div 150 = 1.53$$

The  $z$ -score for a value of 1380 is 1.53. That means 1380 is 1.53 standard deviations from the mean of your distribution.

# What is the central limit theorem in statistics?

The **central limit theorem** states that if you have a population with mean  $\mu$  and standard deviation  $\sigma$  and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally



# What are the three parts of the central limit theorem?

To wrap up, there are **three different components of the central limit theorem**: Successive sampling from a population. Increasing sample size.

...

**Understanding the central limit theorem**

$\mu$  is the population mean.

$\sigma$  is the population standard deviation.

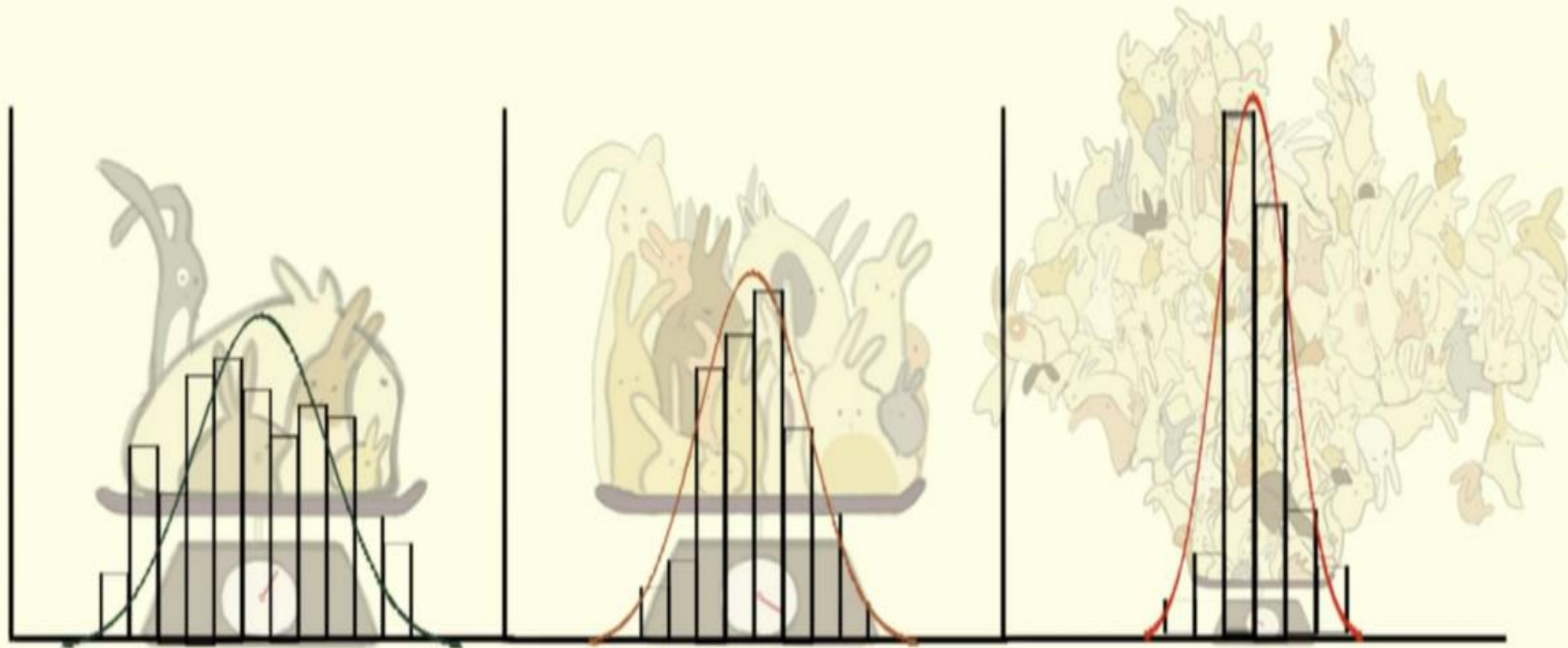
$n$  is the sample size.



# What is central limit theorem and why is it important?

The Central Limit Theorem is important for statistics because it allows us to safely assume that the **sampling distribution** of the mean will be normal in most cases. This means that we can take advantage of statistical techniques that assume a normal distribution, as we will

# Central Limit Theorem



The averages of samples have **approximately normal distributions**

Sample size  $\longrightarrow$  **Bigger**

Distribution of Averages  $\longrightarrow$  **More normal and narrower**



## Practical Applications of CLT

**1 in 3**

Chance Democrats  
win control (34.3%)

**2 in 3**

Chance Republican  
keep control (65.7%)

↑  
Higher  
probability

Breakdown of seats by  
party

Democrat Seats (D)	Republican Seats (R)
55D	45R
54D	46R
53D	47R
52D	48R
51D	49R
50D	50R
49D	51R
48D	52R
47D	53R
46D	54R
45D	55R
44D	56R
43D	57R
42D	58R
41D	59R

CONTROL

CURRENT  
BREAKDOWN

+3

+0.4 Democratic seats  
AVG. GAIN

Source: [projects.fivethirtyeight.com](http://projects.fivethirtyeight.com)

- Political/election polls are prime CLT applications. These polls estimate the percentage of people who support a particular candidate. You might have seen these results on news channels that come with confidence intervals. The central limit theorem helps calculate that
- Confidence interval, an application of CLT, is used to calculate the mean family income for a particular region

The central limit theorem has many applications in different fields. Can you think of more examples? Let me

# Assumptions Behind the Central Limit Theorem

Before we dive into the implementation of the central limit theorem, it's important to understand the assumptions behind this technique:

1. The **data must follow the randomization condition**. It must be sampled randomly
2. **Samples should be independent of each other**. One sample should not influence the other samples
3. **Sample size should be not more than 10% of the population** when sampling is done without replacement
4. The **sample size should be sufficiently large**. Now, how we will figure out how large this size should be?  
Well, it depends on the population. When the population is skewed or asymmetric, the sample size should be large. If the population is symmetric, then we can draw small samples as well

In general, **a sample size of 30 is considered sufficient when the population is symmetric**.

The mean of the sample means is denoted as:

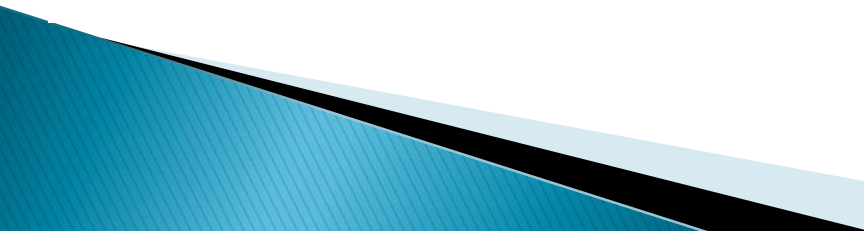
$$\mu_{\bar{x}} = \mu$$

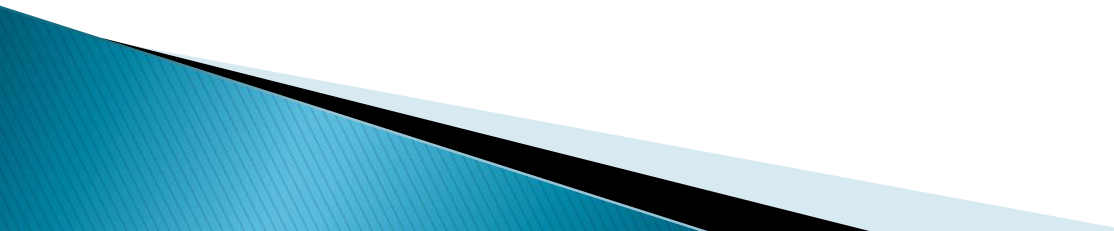
- $\mu_{\bar{x}}$  = Mean of the sample means
- $\mu$  = Population mean

And, the standard deviation of the sample mean is denoted as:

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

where,

- $\sigma_{\bar{x}}$  = Standard deviation of the sample mean
  - $\sigma$  = Population standard deviation
  - $n$  = sample size
- 

- The Central Limit Theorem (CLT) is one of the most popular theorems in statistics and it's very useful in real world problems.
  - In a lot of situations where you use statistics, the ultimate goal is to identify the characteristics of a *population*.
  - Central Limit Theorem is an approximation you can use when the population you're studying is so big, it would take a long time to gather data about each individual that's part of it.
  - Central Limit Theorem helps you balance the time and cost of collecting all the data you need to draw conclusions about the population.
- 


- What Is the Standard Error?

The standard error (SE) of a statistic is the approximate standard deviation of a statistical sample population. The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation. In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error of the mean.

- The **standard error (SE)** of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation. If the statistic is the sample mean, it is called the **standard error of the mean (SEM)**.

- What standard error tells us?

The **standard error tells** you how accurate the **mean** of any given sample from that population is likely to be compared to the true population **mean**. When the **standard error** increases, i.e. the means are more spread out, it becomes more likely that any given **mean** is an inaccurate representation of the true population **mean**.



What is the formula for calculating standard error?

Step 1: **Calculate** the mean (Total of all samples divided by the number of samples).

Step 2: **Calculate** each measurement's **deviation** from the mean (Mean minus the individual measurement).

Step 3: Square each **deviation** from mean. ...

Step 7: Divide the **standard deviation** by the square root of the sample size (n).

## STANDARD ERROR CALCULATION

### Procedure:

- Step 1: Calculate the mean (Total of all samples divided by the number of samples).
- Step 2: Calculate each measurement's deviation from the mean (Mean minus the individual measurement).
- Step 3: Square each deviation from mean. Squared negatives become positive.
- Step 4: Sum the squared deviations (Add up the numbers from step 3).
- Step 5: Divide that sum from step 4 by one less than the sample size ( $n-1$ , that is, the number of measurements minus one)
- Step 6: Take the square root of the number in step 5. That gives you the "standard deviation (S.D.)."
- Step 7: Divide the standard deviation by the square root of the sample size ( $n$ ). That gives you the "standard error".
- Step 8: Subtract the standard error from the mean and record that number. Then add the standard error to the mean and record that number. You have plotted  $\text{mean} \pm 1$  standard error (S. E.), the distance from 1 standard error below the mean to 1 standard error above the mean



**Example:**

Name	Height to nearest 0.5 cm	<b>2</b> Deviations (m-i)	<b>3</b> Squared deviations (m-i) <sup>2</sup>
1. Waldo	150.5	11.9	141.61
2. Finn	170.0	-7.6	57.76
3. Henry	160.0	2.4	5.76
4. Alfie	161.0	1.4	1.96
5. Shane	170.5	-8.1	65.61
n= 5	<b>1</b> Mean m = 162.4 cm		<b>4</b> Sum of squared deviations $\Sigma(m-i)^2 = 272.70$

**5** Divide by number of measurements-1.  $\Sigma (m-i)^2 / (n-1) = 272.70 / 4 = 68.175$

**6** Standard deviation = square root of  $\Sigma (m-i)^2 / n-1 = \sqrt{68.175} = 8.257$

**7** Standard error = Standard deviation/ $\sqrt{n} = 8.257/2.236 = 3.69$

**8**  $m \pm 1SE = 162 \pm 3.7$  or 159cm to 166cm for the men (162.4 - 3.7 to 162.4 + 3.7).

# What is an Estimator?

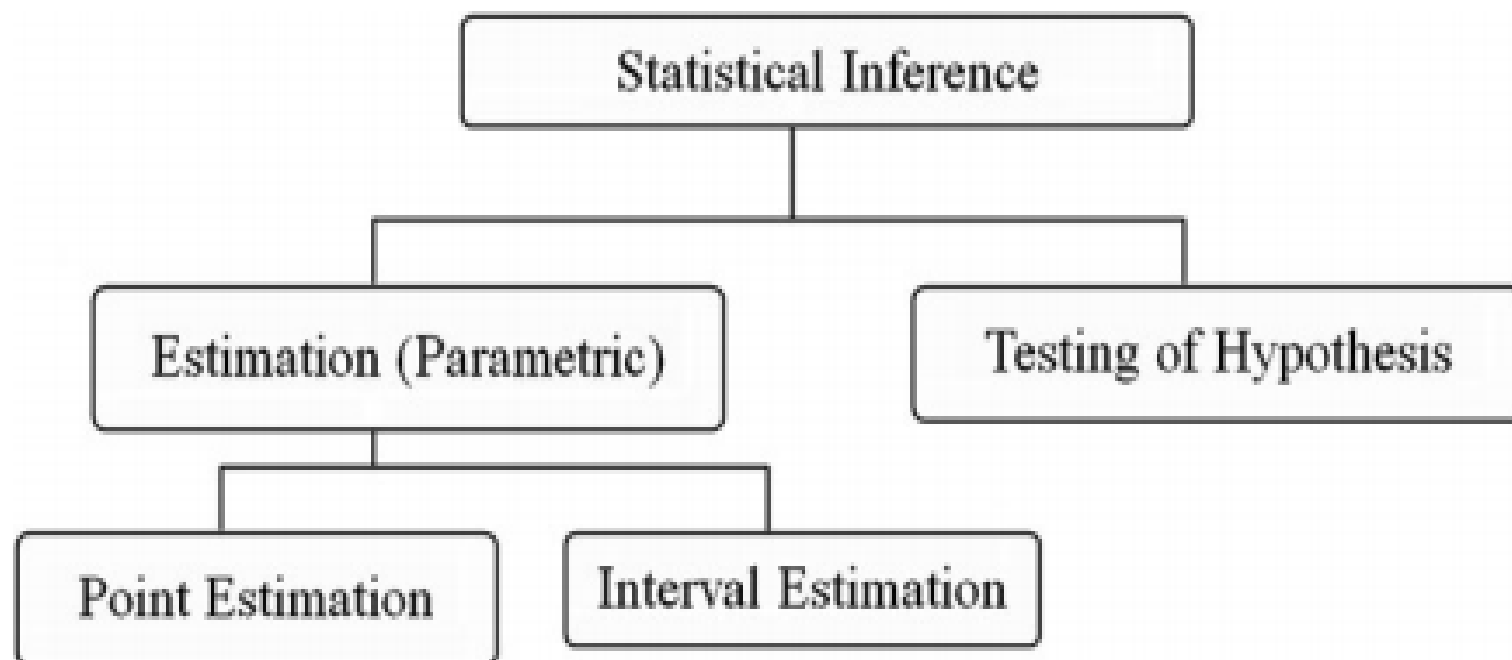
$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i.$$

*The sample mean is  
an estimator for the  
population mean.*

---

An estimator is a **statistic** that estimates some fact about the population. You can also think of an estimator as the rule that creates an estimate. For example, the **sample mean**( $\bar{\mathbf{x}}$ ) is an estimator for the population mean,  $\mu$ .

The quantity that is being estimated (i.e. the one you want to know) is called the **estimand**. For example, let's say you wanted to know the average height of children in a certain school with a population of 1000 students. You take a sample of 30 children, measure them and find that the mean height is 56 inches. This is your sample mean, the **estimator**. You use the sample mean to **estimate** that the population mean (your **estimand**) is about 56 inches.



## Point vs. Interval

Estimators can be a **range of values** (like a **confidence interval**) or a **single value** (like the **standard deviation**). When an estimator is a range of values, it's called an **interval estimate**. For the height example above, you might add on a confidence interval of a couple of inches either way, say 54 to 58 inches. When it is a single value – like 56 inches – it's called a **point estimate**.

## Types

Estimators can be described in several ways (click on the **bold** word for the main article on that term):

**Biased**: a statistic that is either an overestimate or an underestimate.

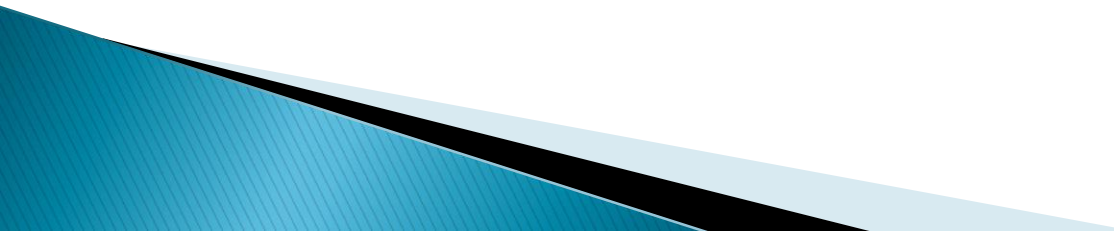
**Efficient**: a statistic with small variances (the one with the smallest possible variance is also called the “best”). *Inefficient* estimators can give you good results as well, but they usually requires much larger samples.

**Invariant**: statistics that are not easily changed by transformations, like simple data shifts.

**Shrinkage**: a raw estimate that’s improved by combining it with other information. See also: **The James–Stein estimator**.

**Sufficient**: a statistic that estimates the population parameter as well as if you knew all of the data in all possible samples.

**Unbiased**: an accurate statistic that neither underestimates nor overestimates.



# Confidence Intervals



In statistics, a **confidence interval (CI)** is a type of estimate computed from the statistics of the observed data. This proposes a range of plausible values for an unknown parameter. The interval has an associated **confidence level** that the true parameter is in the proposed range. This is more clearly stated as: the confidence level represents the probability that the unknown parameter lies in the stated interval. The level of confidence can be chosen by the investigator. In general terms, a confidence interval for an unknown parameter is based on sampling the distribution of a corresponding estimator.

*Imagine you want to find the mean height of all the people in a particular US state.*

You could go to each person in that particular state and ask for their height, or you can do the smarter thing by taking a sample of 1000 people in the state.

Then you can use the mean of their heights (**Estimated Mean**) to estimate the average of heights in the state(**True Mean**)

This is all well and good, but you being the true data scientist, are not satisfied. The estimated mean is just a single number, and you want to have a range where the true mean could lie.

*Why do we want a range? Because in real life, we are concerned about the confidence of our estimates.*

Typically even if I ask you to guess the height of people in the particular US state, you are more inclined to say something like: “*I believe it is between 6 foot to 6 Foot 2 Inch*” rather than a point estimate like “*Its 6 foot 2.2345*”



*inches”.*

We humans also like to attach a level of confidence when we give estimates. Have you ever said — “I am 90% confident”.


In this particular example, I can be more confident about the statement- *“I believe it is between 5 foot to 7 Foot”* than *“I believe it is between 6 foot to 6 Foot 2 Inch”* as the first range is a superset of the second one.

So how do we get this range and quantify a confidence value?

. . .

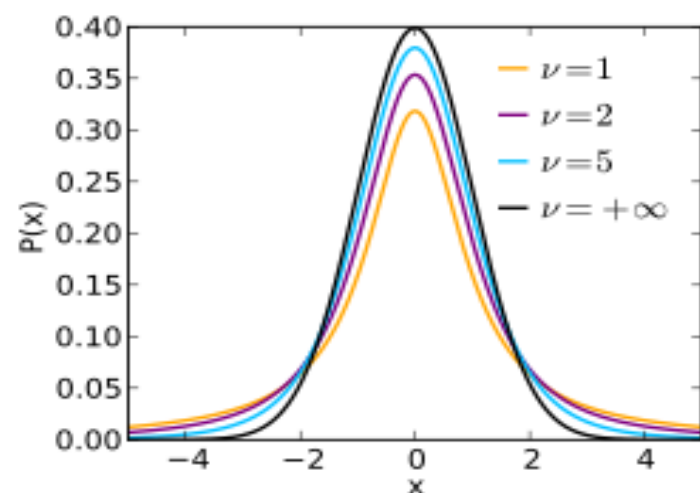
## Strategy

To understand how we will calculate the confidence intervals, we need to understand the Central Limit Theorem.



# What is the T Distribution?

The **T distribution** (also called **Student's T Distribution**) is a family of distributions that look almost identical to the **normal distribution** curve, only a bit shorter and fatter. The t distribution is used instead of the normal distribution when you have small samples (for more on this, see: **t-score vs. z-score**). The larger the **sample size**, the more the t distribution looks like the normal distribution. In fact, for sample sizes larger than 20 (e.g. more degrees of freedom), the distribution is almost exactly like the normal distribution.



## How to Calculate the Score for a T Distribution

When you look at the **t-distribution tables**, you'll see that you need to know the "df." This means "**degrees of freedom**" and is just the sample size minus one.

**Step 1:** Subtract one from your sample size. This will be your degrees of freedom.

**Step 2:** Look up the df in the left hand side of the **t-distribution table**. Locate the column under your **alpha level** (the alpha level is usually given to you in the question).



# What is a Margin of Error?

A **margin of error** tells you **how many percentage points your results will differ** from the real population value. For example, a 95% **confidence interval** with a 4 percent margin of error means that your **statistic** will be within 4 percentage points of the real population value 95% of the time.

More technically, the **margin of error** is the **range** of values below and above the **sample statistic** in a **confidence interval**. The confidence interval is a way to show what the **uncertainty** is with a certain **statistic** (i.e. from a poll or survey).

For example, a poll might state that there is a 98% confidence interval of 4.88 and 5.26. That means if the poll is repeated using the same techniques, 98% of the time the true population parameter (**parameter vs. statistic**) will fall within the interval estimates (i.e. between 4.88 and 5.26) 98% of the time.

