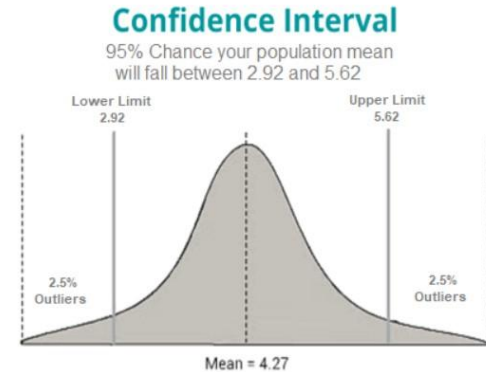




# Measures of Dispersion & Central Limit Theorem

# Agenda - Schedule

1. Warm-Up
2. Measures of Dispersion & Distributions
3. Central Limit Theorem
4. Python Implementations & Plotting
5. TLAB # 1



*Where do we expect most of our data to fall?*



## Agenda - Goals

- Interpret probability distributions
- Understand measures of dispersion
- Identify the importance of the central limit theorem

# Warm-Up

---

**You are working with a dataset of users and the amount of minutes they spend on your shopping platform.**

**The amount of minutes that a user spends is normally distributed, with an average of 4.26 minutes.**

**The standard deviation is 1.28 minutes.**

**One of your users spends 6.91 minutes on your platform. Considering that is is a normal distribution with the stated mean and standard deviation, is this user an outlier? Why or why not?**

Join your pod groups and evaluate this statistics problem. Work together to figure out what will occur when we run this code.

# Stats - Dispersion

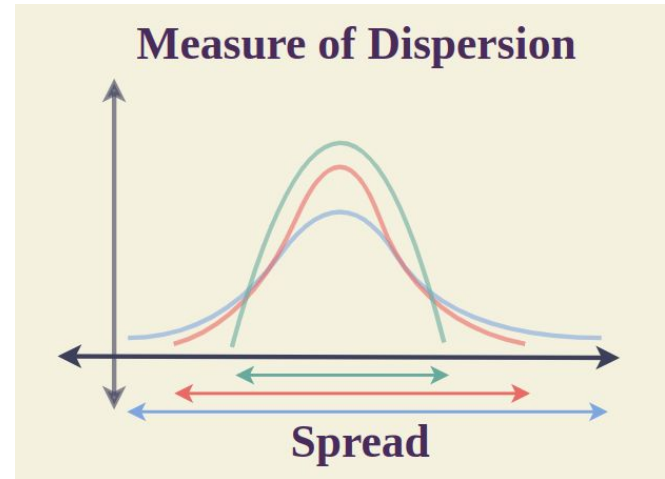
---

## Measures of Dispersion

So far, we've learned about different ways to measure the “**typicality**” of a dataset. This is an important descriptive statistic that allows us to “summarize” typical values.

We also can measure the **variability** of our data. These measures tell us how spread out the data is.

*“How “far” away from the mean or median do the observed values tend to be?” (Learning Statistics with R - Danielle Navarro)*

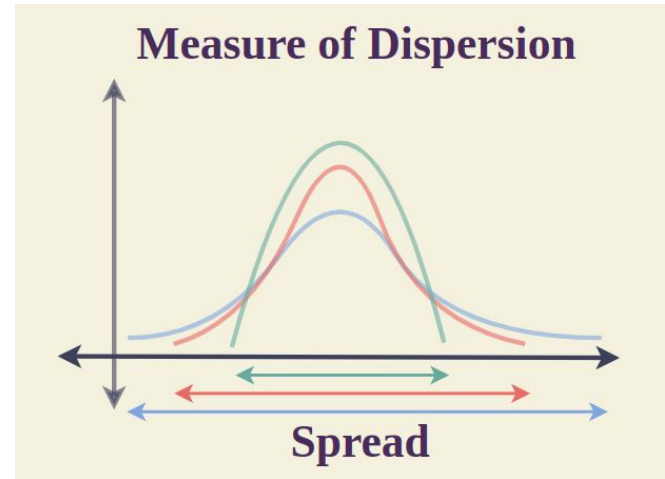


## Measures of Dispersion

To describe our **variance** we have a few options to use, each that have their own unique benefits:

- *Range*
- *Interquartile Range (IQR)*
- *Mean absolute deviation (MAD)*
- *Variance ( $\sigma^2$ )*
- *Standard Deviation ( $\sigma$ )*

Before we explore each definition, let's determine some common mathematical notation that are vital for these concepts.





Symbol	Meaning	Refers to a value in a
$n$	Sample size	Sample
$\bar{x}$	Sample mean	Sample
$\mu$	Population mean	Population
$s$	Sample standard deviation	Sample
$\sigma$	Population standard deviation	Population

We will not discuss the relationship between a “sample” and “population” just yet, but just keep in mind that the sample is an “estimate” of the population.

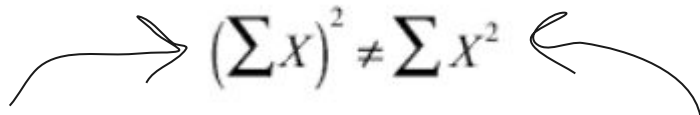
# Summation

We'll run into a lot of summation in statistics, the act of summing up numbers

Here are some key things to note:

$\sum X$  This symbol means “*Sum all X where X is a variable from your dataset.*”

Please also note:


$$(\sum X)^2 \neq \sum X^2$$

Sum up all the values of X and square the sum

Square each value of X, and then sum

	<b>X</b>	<b>X<sup>2</sup></b>
	1	1
	3	9
	4	16
	5	25
	6	36
<b>ΣX</b>	?	
<b>(ΣX)<sup>2</sup></b>	?	
<b>ΣX<sup>2</sup></b>		?

Just to denote the difference between these two calculations, let's calculate these values by hand...

	<b>X</b>	<b>X<sup>2</sup></b>
	1	1
	3	9
	4	16
	5	25
	6	36
<b>ΣX</b>	19	
<b>(ΣX)<sup>2</sup></b>	391	
<b>ΣX<sup>2</sup></b>		87

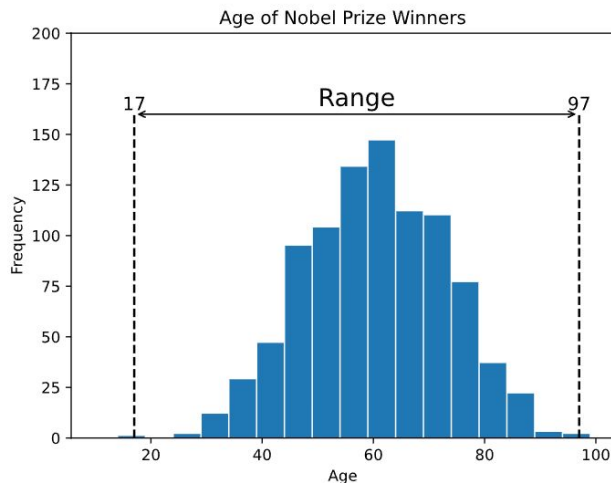
Just to denote the difference between these two calculations, let's calculate these values by hand...

## Measures of Dispersion - Range

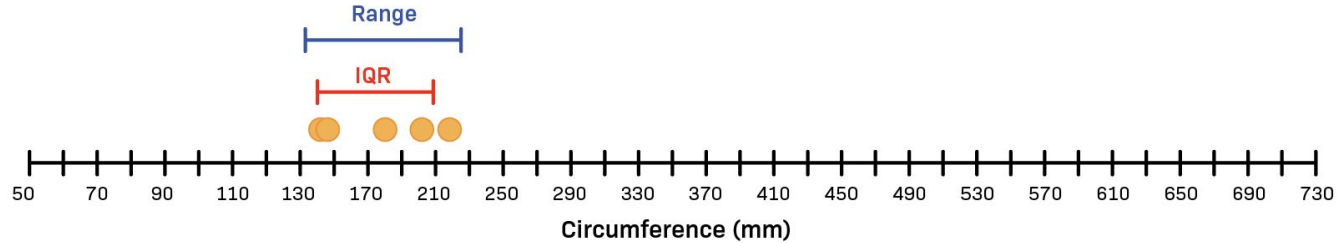
The first concept we will explore is the **range**.

This is a naive calculation, and simply entails pulling the minimum and the maximum value from the dataset and calculating the **difference**.

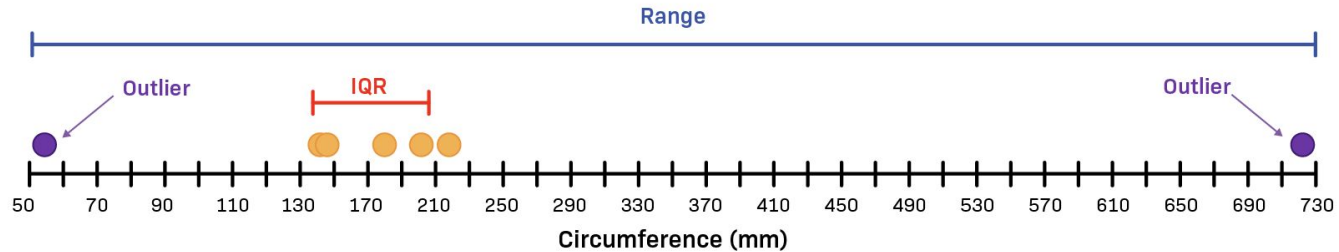
While easy to calculate, this is also the **worst measure of variability** as could give us an inaccurate portrayal of the data if outliers exist.



A. **IQR** in the *absence* of an **outlier**. Orange trees with circumferences of **140**, **145**, **177**, **203**, and **214** mm.



B. **IQR** in the *presence* of **outliers**. Orange trees with circumferences of **52**, **140**, **145**, **177**, **203**, **214**, and **700** mm.



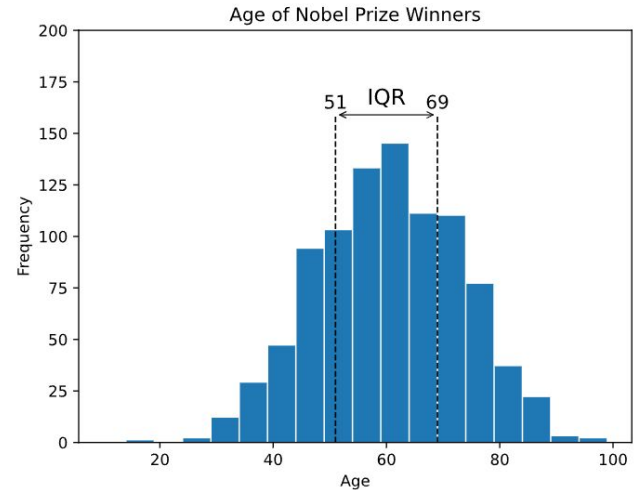
Let's say most of our data is clustered around 180. If this data does not suffer from outliers, our range will be an accurate portrayal of the variability. However, **when we introduce outliers our range quickly blows up.**

## Measures of Dispersion - IQR

To account for the non-robustness of range, we utilize the **interquartile-range (IQR)** of a dataset.

Here we calculate the 25th **quantile** and 75th **quantile** of the dataset and subtract these two values.

A 25th **quantile** (or **percentile**) is the smallest number  $x$  such that 25% of the data is less than  $x$ .



```
data = [-100, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100]
```

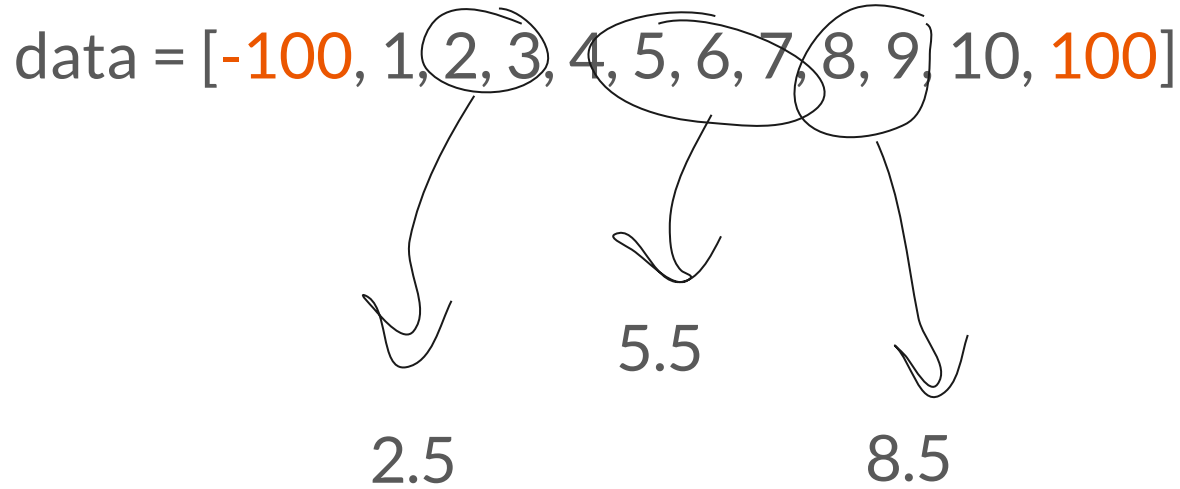
What we will get if we calculate the range of this dataset?



data = [-100, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100]

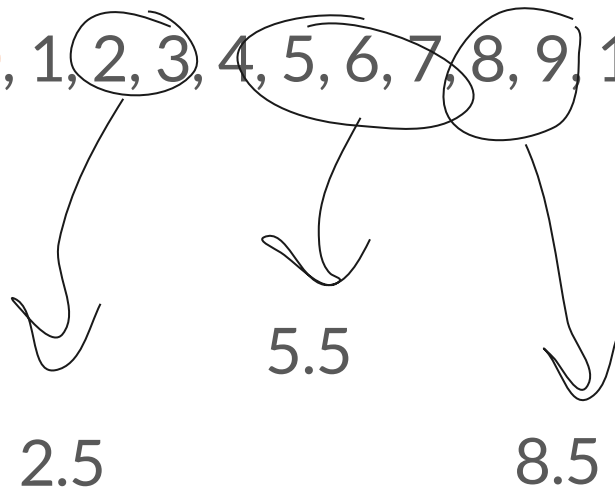
range = 100 - -100 = 200

This is undoubtedly an inaccurate measure that fails to capture the “closeness” of this dataset.



So instead, we calculate the IQR. First we get the median of this dataset. From there we **calculate the median of the two groups that are subsequently formed on both sides of the median.**

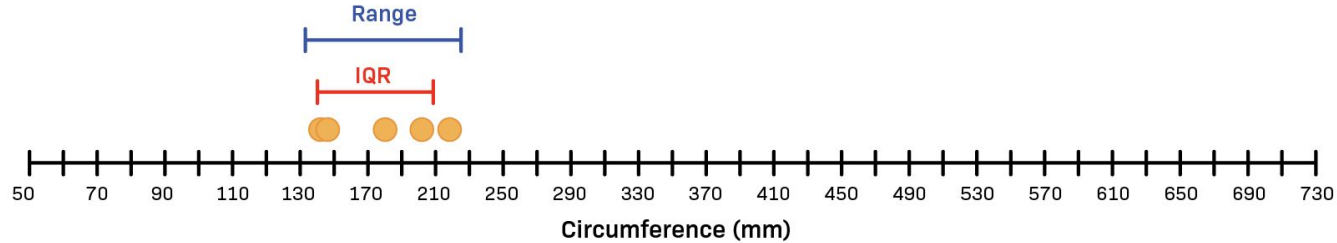
data = [-100, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100]



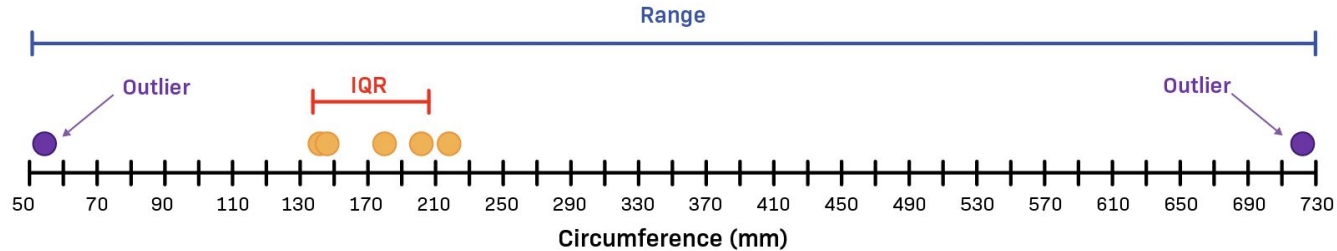
$$8.5 - 2.5 = 6$$

We then calculate the difference between the two quartiles. This gives us our IQR, which is a much more accurate measure of variability than range gave us.

A. **IQR** in the *absence* of an **outlier**. Orange trees with circumferences of **140**, **145**, **177**, **203**, and **214** mm.



B. **IQR** in the *presence* of **outliers**. Orange trees with circumferences of **52**, **140**, **145**, **177**, **203**, **214**, and **700** mm.



Let's say most of our data is clustered around 180. If this data does not suffer from outliers, our range will be an accurate portrayal of the variability. However, **when we introduce outliers our range quickly blows up.**



## Measures of Dispersion - Mean Absolute Deviation

*A different approach is to select a meaningful reference point (usually the mean or the median) and then report the “typical” deviations from that reference point.*

While it is possible to use median as your reference point, we will instead opt to use the **mean** (since this is more common).

The name is self-explanatory, calculate all the absolute differences between the mean and each data point, sum, and divide by **n**.

### Mean Deviation

$$MD = \frac{\sum |x - \bar{x}|}{n}$$

$$MD = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + |x_3 - \bar{x}|}{n}$$

```
data = [5, 6, 8, 9, 10, 16]
```

Let's try this calculation out on this example dataset...

`data = [5, 6, 8, 9, 10, 16]`

`x_bar = 9`

`MAD = (abs(5-9) + abs(6-9) + abs(8-9) +  
abs(10-9) + abs(16-9)) / 6`

`MAD = 2.6`



## Measures of Dispersion - Variance

*From a purely mathematical perspective, there are some solid reasons to prefer squared deviations rather than absolute deviations*

*As you can see, it's basically the same formula that we used to calculate the mean absolute deviation, except that instead of using "absolute deviations" we use "squared deviations".*

**This has excellent additive properties which will be super useful in our predictive models** ( $Z = X + Y \rightarrow \text{Var}(Z) = \text{Var}(X) + \text{Var}(Y)$ )

$$\sigma^2 = \frac{\sum (xi - \bar{x})^2}{N}$$



```
data = [5, 6, 8, 9, 10, 16]
```

Let's try this calculation out on this example dataset...

`data = [5, 6, 8, 9, 10, 16]`

`x_bar = 9`

`Variance = ((5-9)**2 + (6-9)**2 + (8-9)**2  
+(10-9)**2 + (16-9)**2) / 6`

`Variance = 12.66`



## Measures of Dispersion - Standard Deviation

One main issue with variance is the **interpretation of this metric**. Usually, we calculate some very large value that represents our **units<sup>2</sup>**.

To account for this, we take the square root of our variance and **calculate standard deviation**.

This gives us a metric in our original units which we can use to describe our dataset.

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

`data = [5, 6, 8, 9, 10, 16]`

`x_bar = 9`

`Variance = ((5-9)**2 + (6-9)**2 + (8-9)**2  
+(10-9)**2 + (16-9)**2) / 6`

`Variance = 12.66`

`std_dev = 12.66 ** (1/2) = 3.55`

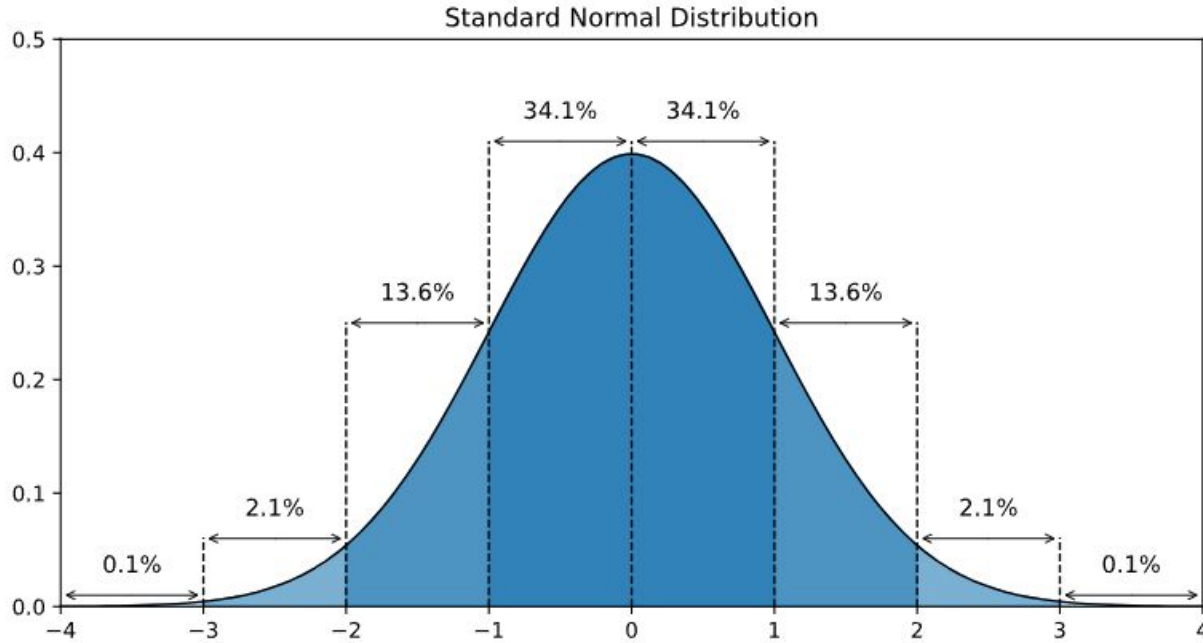
data = [5, 6, 8, 9, 10, 16]

std\_dev =  $12.66^{**}(\frac{1}{2}) = 3.55$

$9 + 3.55 = 12.55$

$9 - 3.55 = 5.45$

When considering our mean and standard deviation, we can see that almost all of the data falls within one standard deviation of the mean. (5.45, 12.55)



This becomes even more applicable when describing the normal distribution. Roughly 68% of your data will fall underneath 1 standard deviation of the mean.

# Probability Distributions

---

# Probability Distributions

Let's go back to our probability distribution.

Truth be told, the previous image we provided is not indicative of how distributions usually look.

First off, we almost always discuss the distributions of a **quantitative variable**.

This means instead of a bar chart, we use a **histogram**.



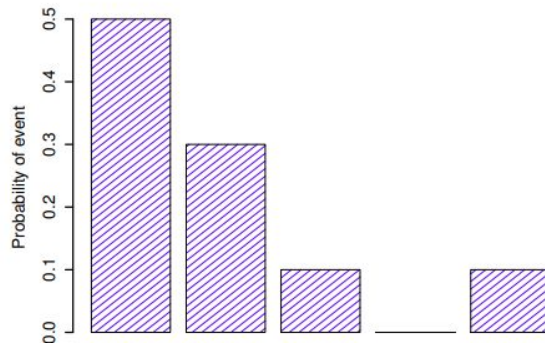


## Probability Distributions

Going back to our discussion of probability theory, we discussed how we can visually represent the **likelihood** of an event occurring by visualizing it using a **bar-chart** (discrete) or **histogram** (continuous).

Simply put, we measure the ratio of “successes” in relation to an event, and plot this using one of the two graphs above.

Which pants?	Label	Probability
Blue jeans	$X_1$	$P(X_1) = .5$
Grey jeans	$X_2$	$P(X_2) = .3$
Black jeans	$X_3$	$P(X_3) = .1$
Black suit	$X_4$	$P(X_4) = 0$
Blue tracksuit	$X_5$	$P(X_5) = .1$

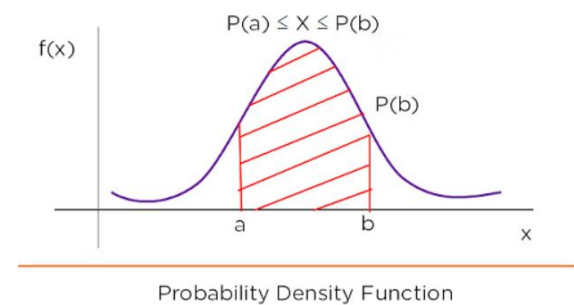


## Probability Density Function (PDF)

When it comes to expressing the probability that a continuous value is **greater than**, **less than**, or **within some range of values** within a random variable, **probability distributions are insufficient.**

Instead of discrete bins, we need a **smooth continuous line** that allows us to express probability as **area under a curve.**

We therefore rely on a tool called the **probability density function** which describes this continuous curve.



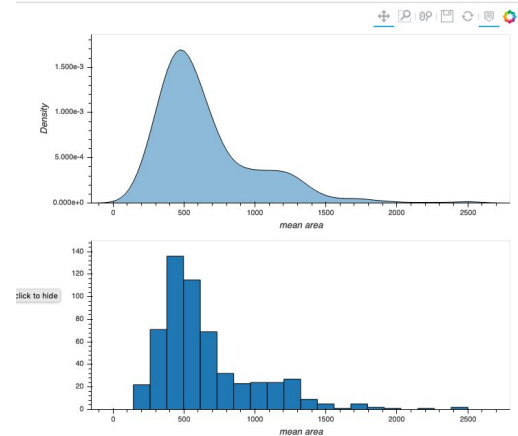
# Probability Density Function (PDF)

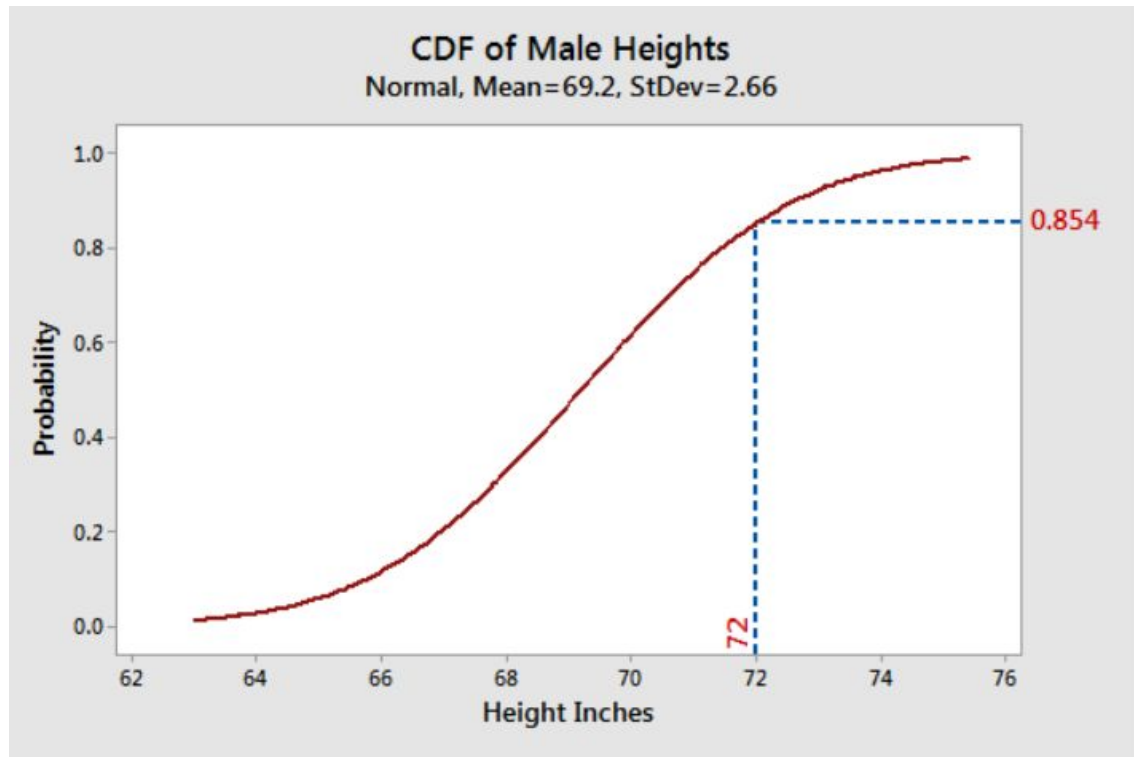
When we plot histograms we have a discrete number of bins that express the frequency of some range of quantities so how do we get the PDF?

First we calculate the cumulative distribution function which expresses the probability that a random variable  $X$  is less than or equal to some value  $x$ .

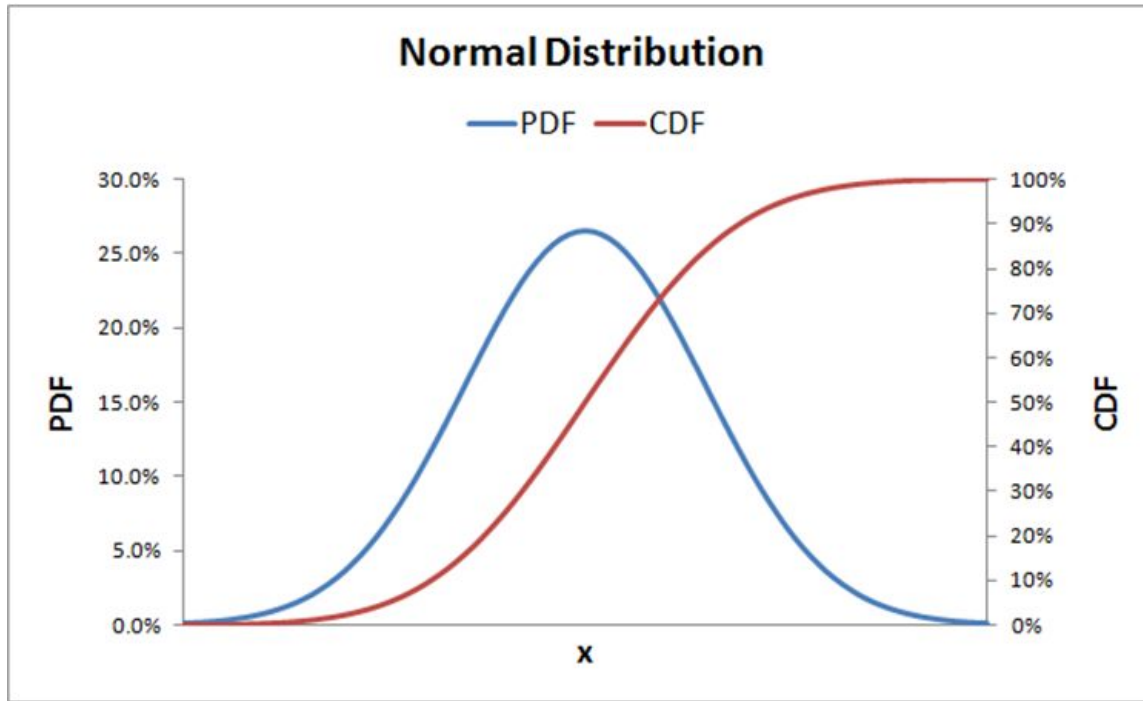
We express this as a function “F”

$$F(x) = P[X \leq x]$$





For example, let's say we have a dataset of **male US heights**. The **CDF** of this dataset would express a graph that displays the **probabilities that male height is less than or equal to a range of possible heights**. Since we are considering all possible values (not just real values), this gives us a **continuous curve**.



$$f_X(x) = \frac{dF_X(x)}{dx}$$

By taking the **derivative** of this curve (*which we will explore in greater detail during phase 2*), we calculate the **probability density function**. This gives us a function where the area under the curve is equal to 1 (*all possible probabilities must sum up to 100%*) and furthermore tells us the behavior of this dataset.

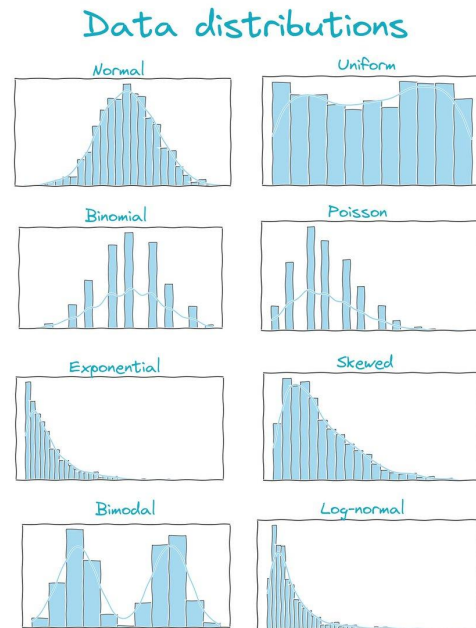
# Probability Distributions - Types of Distributions

When we observe different phenomena, we will observe equally different PDF's.

Let's review the following density functions:

- Binomial Distribution:
- Normal Distribution
- Uniform distribution
- T-distribution
- Chi-Squared distribution

*These are described as parametric models, which indicate that they are modified by some parameters we may or may not know*





## Probability Distributions - Binomial Distribution

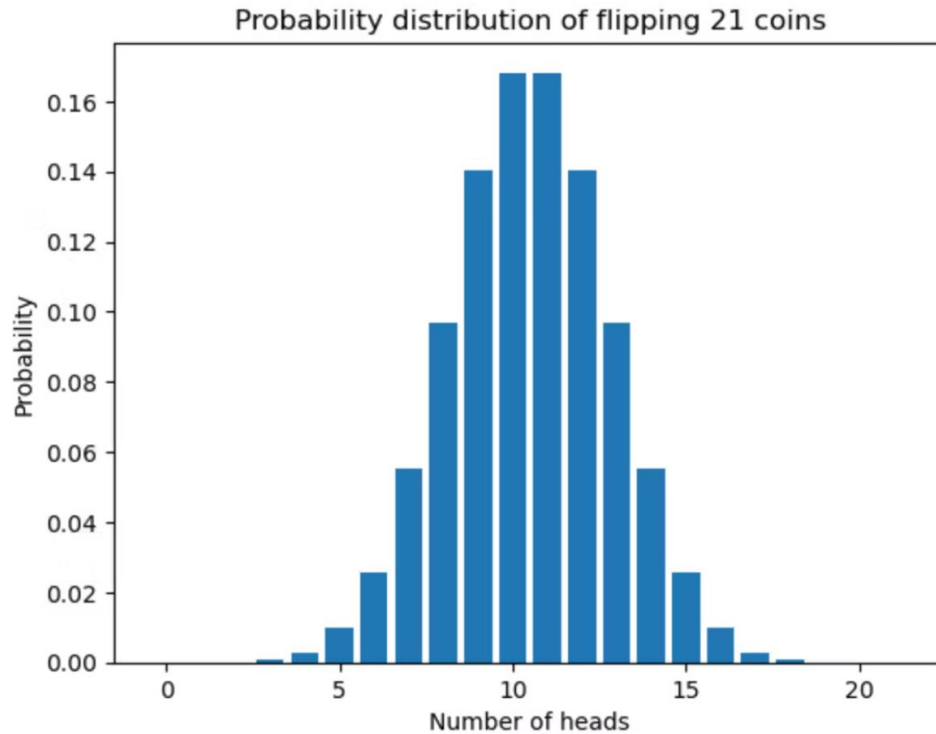
We see the binomial distribution to describe the likelihood of success in repeated games of chance. The parameters are the following:

**N (size parameter):** Number of trials

**Theta (success probability):** Probability of success

**X (random variable):** The number of successes I am looking for

$$P(X | \theta, N) = \frac{\text{Binomial } N!}{X!(N-X)!} \theta^X (1-\theta)^{N-X}$$



$$P(X | \theta, N) = \frac{\text{Binomial } N!}{X!(N - X)!} \theta^X (1 - \theta)^{N-X}$$

Using this distribution, we can answer questions like “what is the probability we flip more than 19 Heads?” or “what is the probability we flip at least 5 Heads”

Here we calculate the binomial distribution of the **number of heads** we flip in **21 trials** given a **50% chance of flipping heads** (fair coin). While the histogram on the left expresses the experimental data, we can also simply use the binomial distribution formula.





## Probability Distributions - Normal Distribution

We see the normal distribution when observing measurements on real world quantities (*height, miles driven, mosquito wingspan, number of gun crimes in the US*).

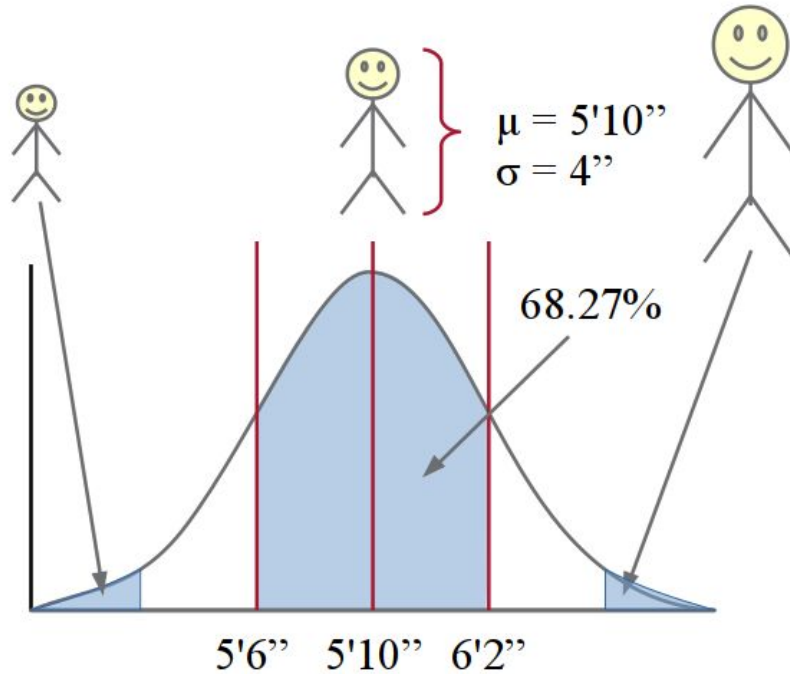
The parameters are:

**mu (mean):** Mean

**std-dev (success probability):** Standard deviation

**X (random variable):** The measurement I am looking for

$$p(X \mid \mu, \sigma) = \frac{\text{Normal}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X - \mu)^2}{2\sigma^2}\right)$$



$$p(X \mid \mu, \sigma) = \frac{\text{Normal}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X - \mu)^2}{2\sigma^2}\right)$$

Using this distribution, we can answer questions like “what is probability that someone is less than 5'10”

Here we calculate the normal distribution of male height where the mean is 65 inches, and the std-dev is 4 inches. Another nice feature of the normal distribution is the assumption that almost all data falls within 3 std-devs of the mean.



## Probability Distributions - Uniform Distribution

We see the uniform distribution to describe the likelihood of **purely** random events. The parameters for this distribution are:

**a:** Maximum of event  $X$

**b:** Minimum of event  $X$

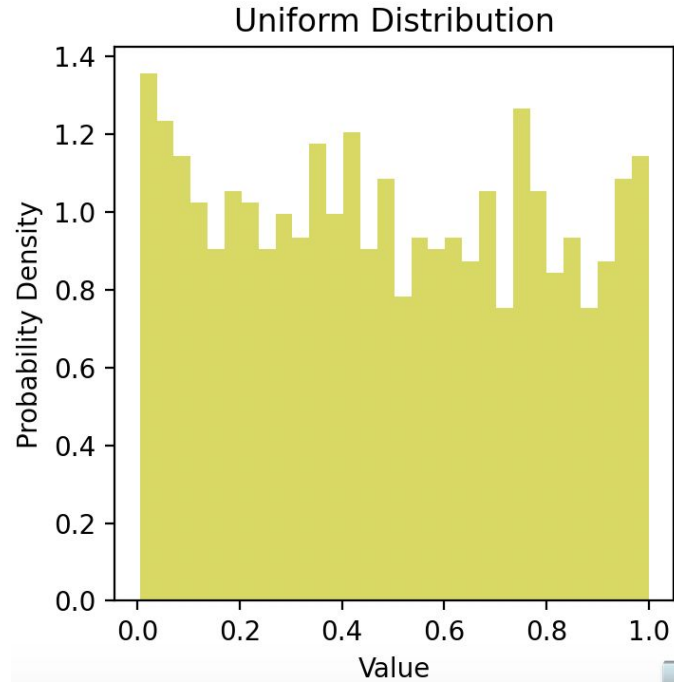
Calculating probabilities of uniform distribution is naive, as it usually just entails calculating elementary probabilities (*prob of 1 coin flip*).

### Uniform Distribution

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

$$\text{Mean } \mu = \frac{a+b}{2}$$

$$\text{Variance } \sigma^2 = \frac{(b-a)^2}{12}$$



The one interesting thing we can point out is that **real-world** random processes **rarely generate a flat and uniform line as we idealize.** The real world is noisy (and we need a lot of trials to get smooth numbers), so you will most likely see uniform distributions with “bumps and ridges.”

# Central Limit Theorem

---



## Central Limit Theorem

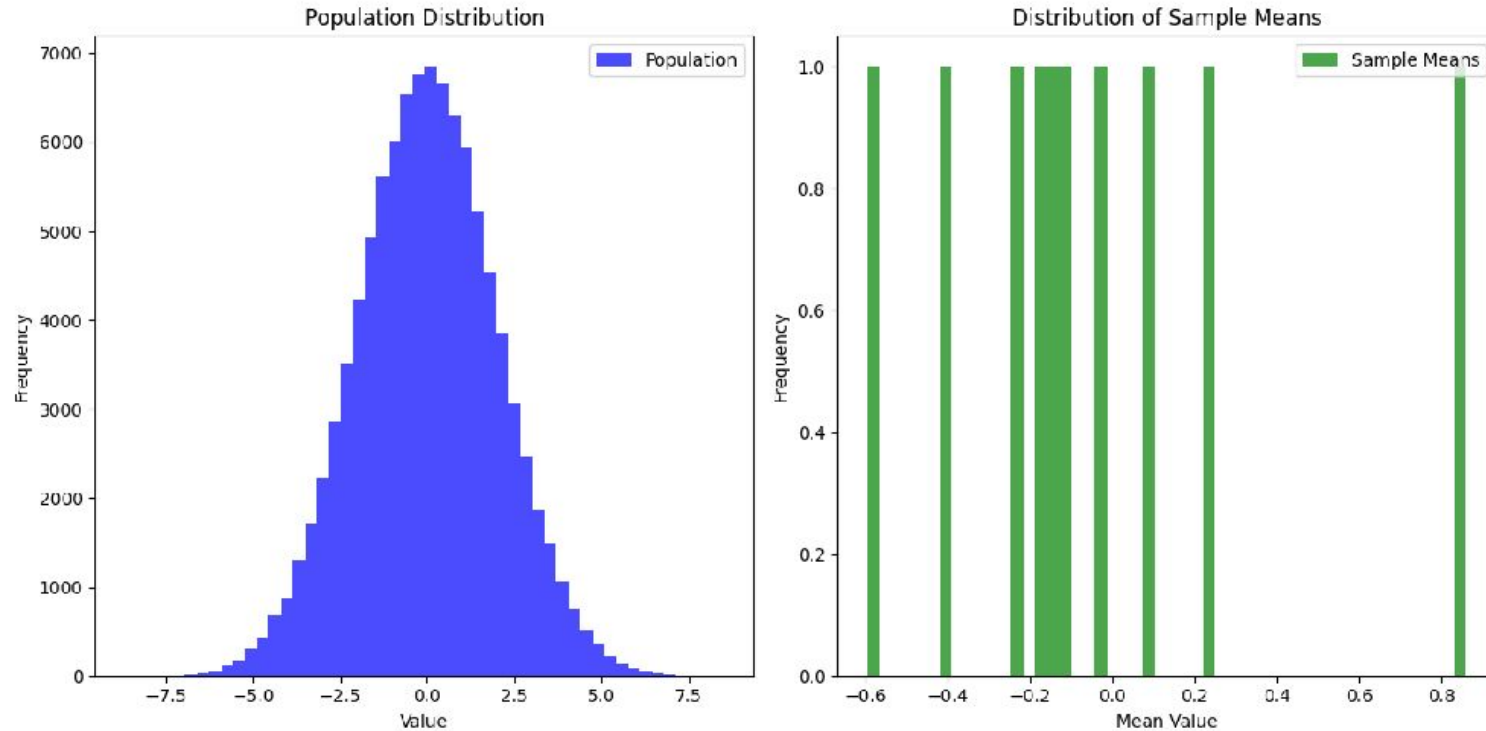
The law of large numbers is a powerful tool, but rather useless. All it tells us is “*as you run your experiments to infinity, the ground truth reveals itself.*”

John Maynard Keynes had a quote on the **long-run**:

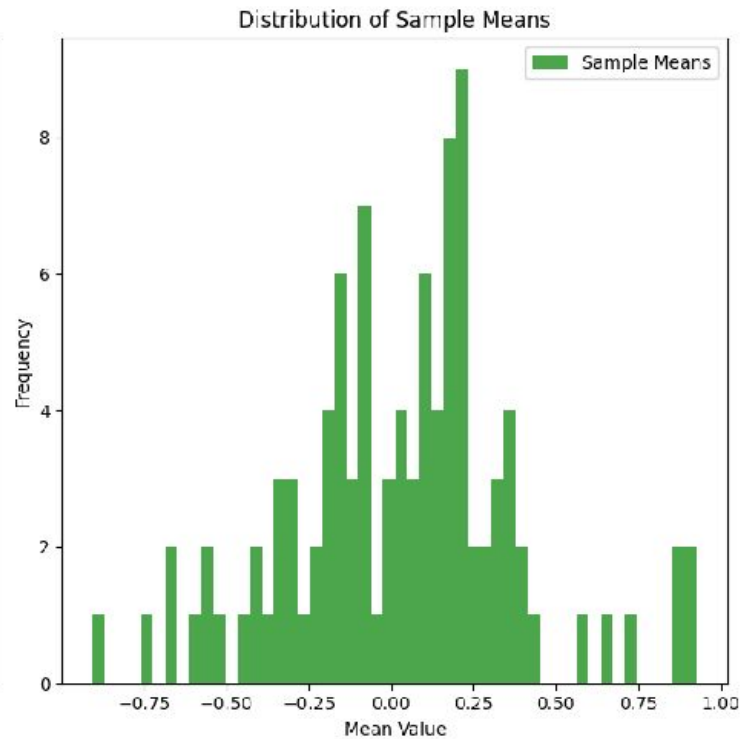
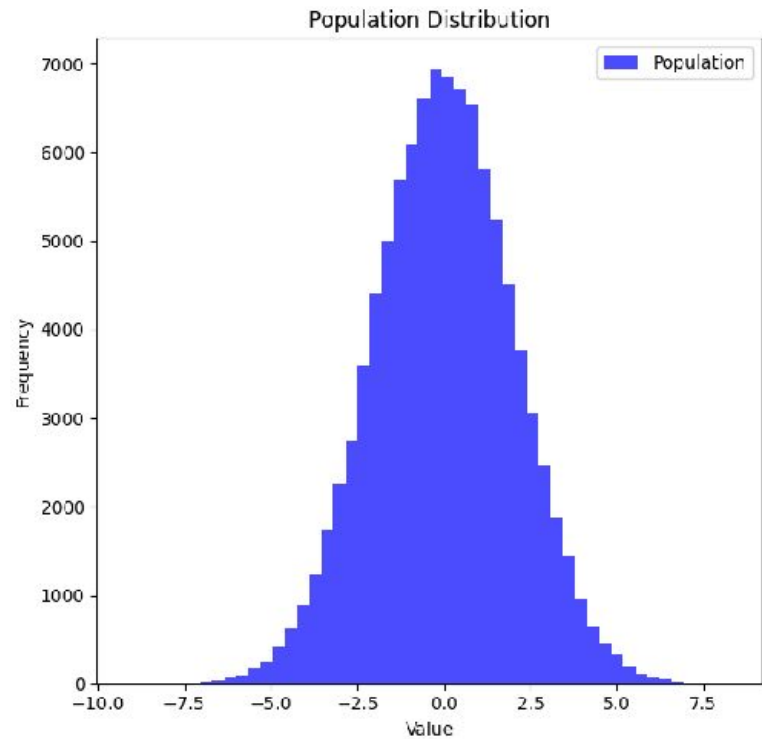
*“[The] long run is a misleading guide to current affairs. In the long run we are all dead...”*

We need a better tool to understand the accuracy of a sample in the context of our population.

First, let's **understand what happens** when we collect samples multiple times and **calculate the mean of our samples in the context of a population.**

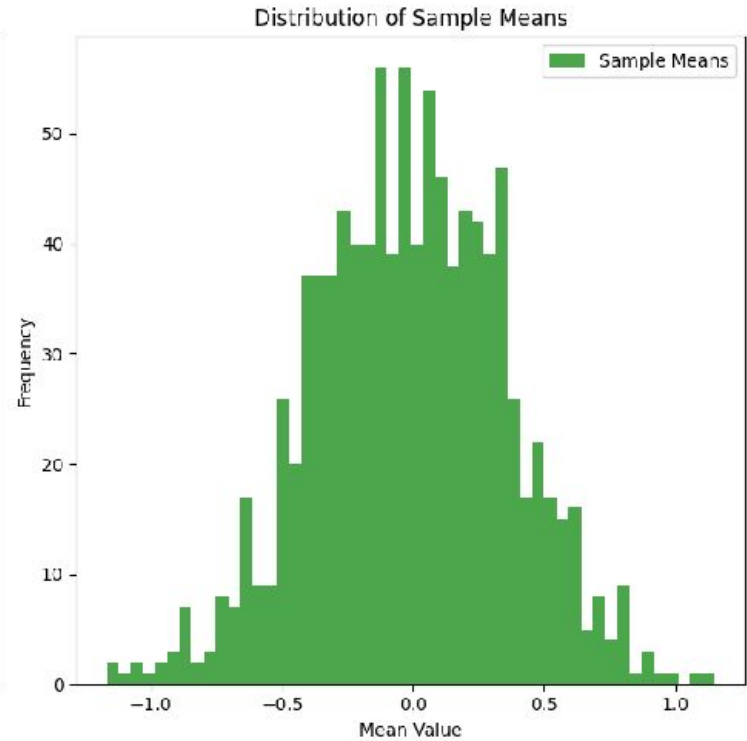
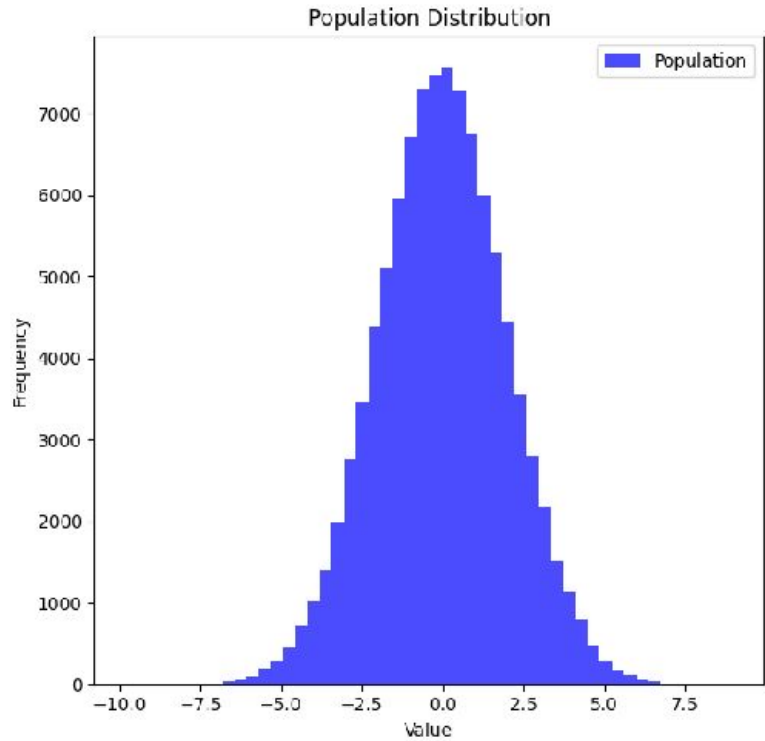


Let's iteratively collect samples on a population (on the left) and calculate the probability distribution of sample means (on the right). **First we start with 10 samples.**



100 samples





1000 samples. What kind of distribution are we forming as we increase the number of samples?

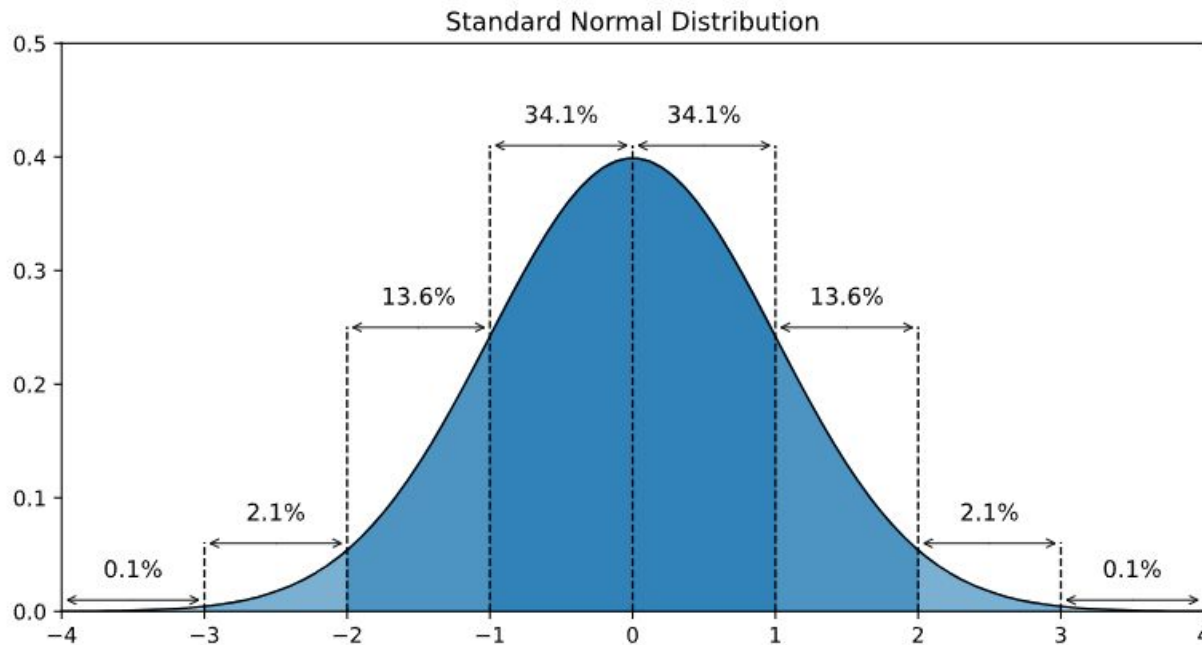


## Central Limit Theorem

As we continue to collect **more** samples on our population and calculate **their sample means**, we start to form a **normal distribution around the original population mean**.

Let's identify why this is such a big deal:

In our **normal distribution**, which percent of data falls within 3 standard deviation of the mean.



This means that roughly any sample that you take from your population (as long as its of size 30, and randomly collected), will be a “good enough” estimate of your population.

Roughly 99% of your data will fall underneath **3 standard deviations** of the mean. In the context of the problem we just explore: **99% of your sample means will fall within 3 standard deviations of the population mean.**

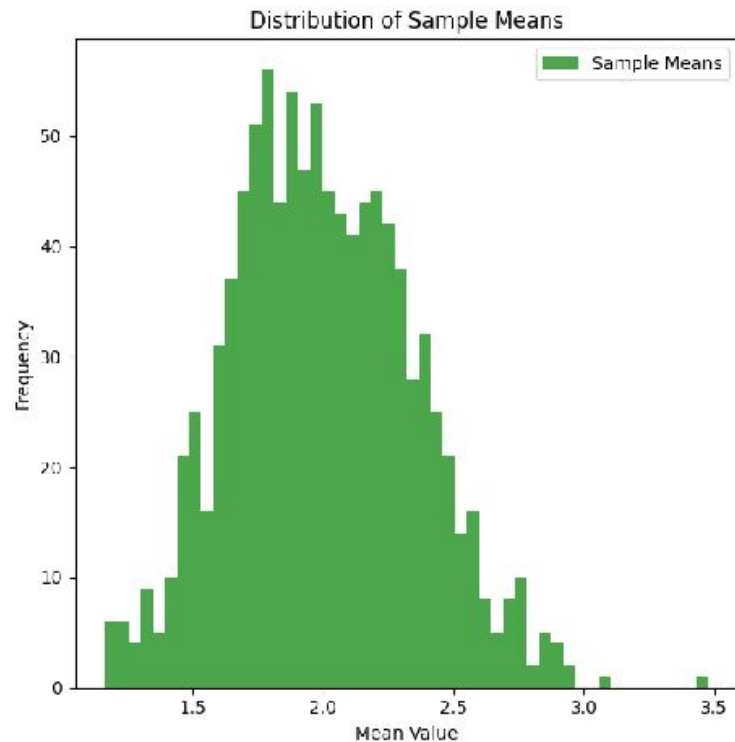
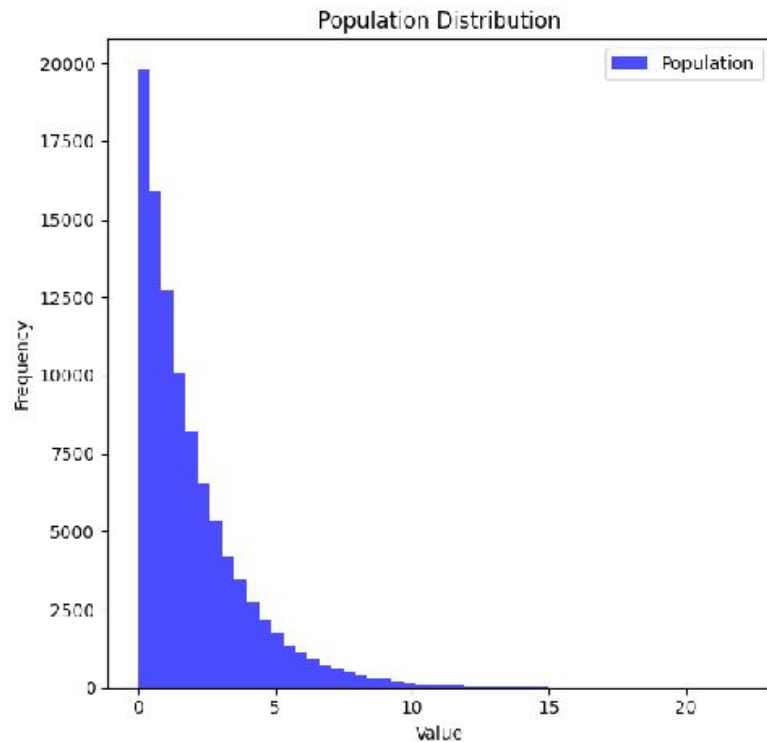


## Central Limit Theorem - Formal Definition

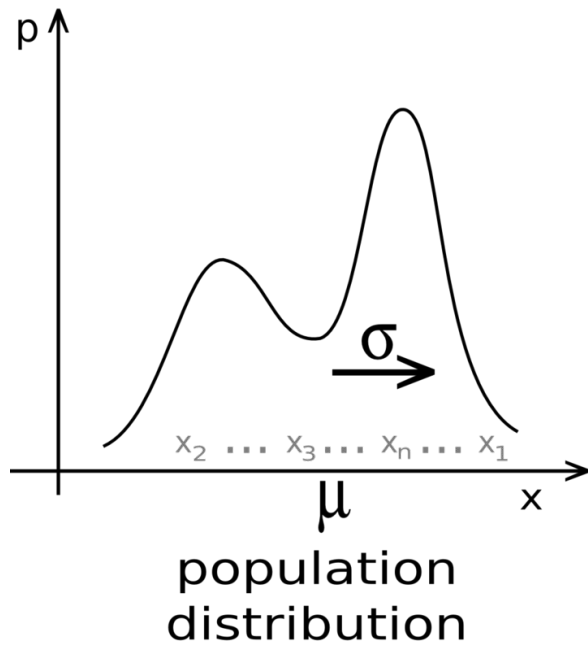
Formally: “...for *independent and identically distributed random variables*, the *sampling distribution of the standardized sample mean tends towards the standard normal distribution* even *if the original variables themselves are not normally distributed.*”

This last part is an even **more ground-break revelation**.

Let's say we have some non-normally distributed population.



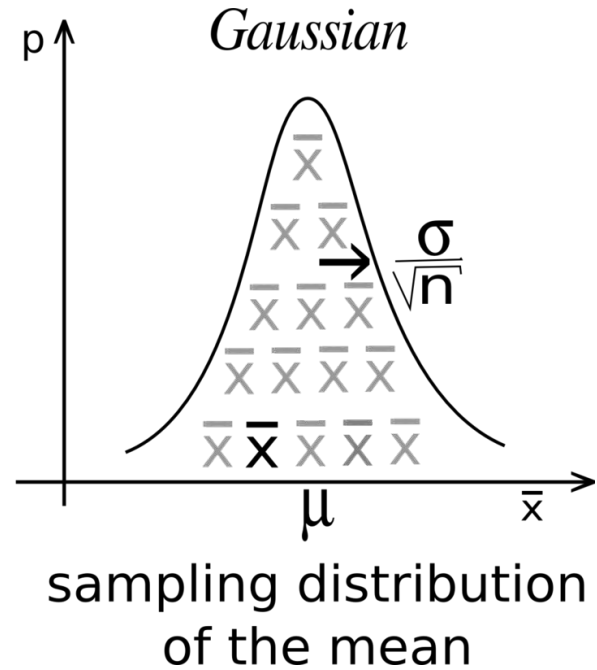
On the left we have an exponential distribution. On the right, we've calculated distribution of 1000 sample means. Have we formed a normal distribution around the mean???



samples  
of size  $n$

$\bar{x}$

$\bar{x}$



This is a powerful idea which allows us to perform machine learning, even on non-normally distributed populations.



# Central Limit Theorem

The CLT relies on the ***law of large numbers*** which says that the bigger the sample, the sample mean gets closer to the population mean

This is important because that means our multiple samples must be a minimum size

This lets us make the assumption that the sample mean we calculate must at least partially reflect the population

So what is the minimum sample size?



## Central Limit Theorem

The CLT relies on the ***law of large numbers*** which says that the bigger the sample, the sample mean gets closer to the population mean

This is important because that means our multiple samples must be a minimum size

This lets us make the assumption that the sample mean we calculate must at least partially reflect the population

So what is the minimum sample size? **30**



# Python Implementations

---



## Python Implementations

Now that we've gone over these various mathematical formulae, we should also understand how we can **implement these concepts** in Python.

Remember, all of these formula could be implemented by importing a package.

However, you will not truly understand an equation until you implement it yourself.



## Python Implementations

For that reason, we will go through each measure of dispersion and explore which Python code we should write.

- Range
- IQR
- Variance
- Standard deviation

Let's begin with range.

[1.2, 2.4, 3.1, 4.7, 5.0, 5.9, 6.3, 7.2, 8.8, 9.1]

Assume we have the following list. How would we calculate **range**?

[1.2, 2.4, 3.1, 4.7, 5.0, 5.9, 6.3, 7.2, 8.8, 9.1]

How about IQR?

[1.2, 2.4, 3.1, 4.7, 5.0, 5.9, 6.3, 7.2, 8.8, 9.1]

Variance?

[1.2, 2.4, 3.1, 4.7, 5.0, 5.9, 6.3, 7.2, 8.8, 9.1]

And lastly, standard deviation.

# Lab - Work on TLAB #1

---



## Lab (Due 03/28)



*Taipei City, Taiwan*

The company you work for, Seng-Links, aims to identify periods when a user sleeps or exercises using their varying recorded heart rates.

Your company has provided you a data folder (*data/*) of **4 files** that contain heart-rate samples from a participant. The participants device records heart rate data every 5 minutes (aka *sampling rate*).

You are tasked with writing code that **processes each data file**. You will utilize test-driven development in order to complete this project.



# Announcements

A couple of assignment-related announcements:

- **Week 2 Pre-Class Quiz**
  - Cohort B: due 3/22
  - Cohort A: due 3/19
- **OOP Quiz (both cohorts due 3/22)**
- **Thursday review session: everyone join Cohort A Zoom Session**



## Thursday

### TLAB 1 Review

- Bring questions
- Work with your peers
- Work outside of class as well

*If you understand what you're doing, you're not learning anything. - Anonymous*



*Jupyter: scratchpad of the data scientist*