Data Science/Analytics Basics

Aniket Chhabra

Welcome to Stats Refresher Session…

"There are three types of lies -- lies, damn lies, and statistics."

"Facts speak louder than statistics"

"All Statistical models are wrong…. Some models are useful"

"99 percent of all statistics only tell 49 percent of the story"

"I can prove anything by statistics except the truth.."

"Facts are stubborn, but statistics are more pliable.."

"Statistics are used much like a drunk uses a lamppost: for support, not illumination"

Statistics...Probability....And....!!!

- A statistics professor plans to travel to a conference by plane. When he passes the security check, they discover a bomb in his carry-on-baggage. Of course, he is hauled off immediately for interrogation.
- "I don't understand it!" the interrogating officer exclaims. "You're an accomplished professional, a caring family man, a pillar of your parish - and now you want to destroy that all by blowing up an airplane!"
- "Sorry", the professor interrupts him. "I had never intended to blow up the plane."
- "So, for what reason else did you try to bring a bomb on board?!"
- "Let me explain. Statistics shows that the probability of a bomb being on an airplane is 1/1000. That's quite high if you think about it - so high that I wouldn't have any peace of mind on a flight."
- "And what does this have to do with you bringing a bomb on board of a plane?"
- "You see, since the probability of one bomb being on my plane is 1/1000, the chance that there are two bombs is 1/1000000. If I already bring one, the chance of another bomb being around is actually 1/1000000, and I am much safer..."

We are Thankful to….

Courseera.org – Data Analysis and Statistical Inference (Dr. Mine Çetinkaya – Rundel – Duke University)

Basic Econometrics, Fourth Edition – Dr. Damodar Gujarati

Fundamentals of Mathematical Statistics – S.C. Gupta and V.K.Kapoor

Fundamentals of Statistics – Goon A M, Gupta M K, B. Dasgupta

Statistical Methods – N.G.Das

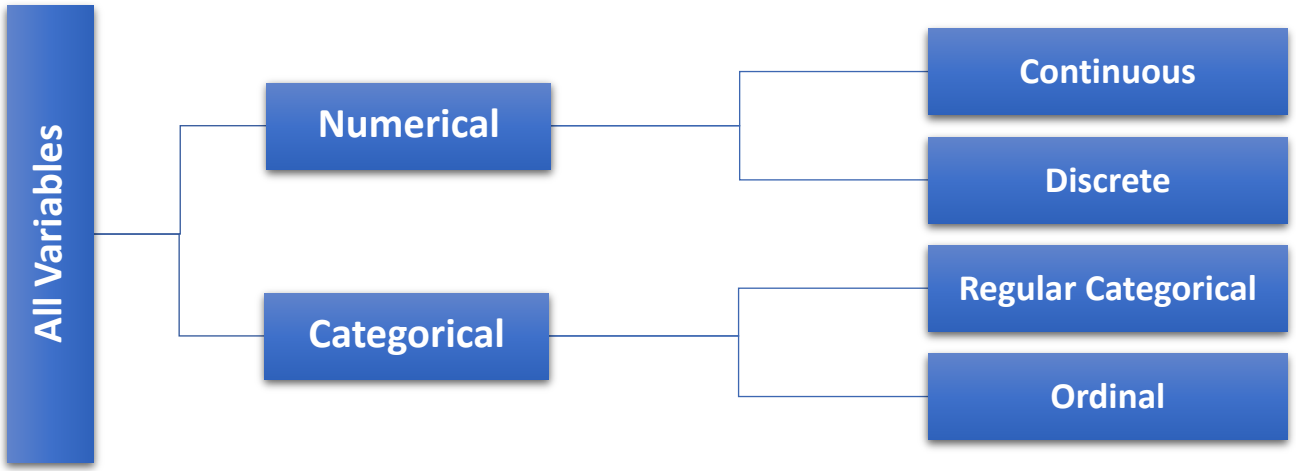
All our Professors and Teachers and Friends and colleagues…

Last But Not Least – the Internet

Course Scope….

- Data Basic
- Sampling and Sources of bias
- Central Tendency
- Measure of Spread
- Visualizing Numeric Data
- Data Transformation
- Exploring Categorical Variables
- Concept of Probability
- Normal Distribution, Binomial and Poison Distribution
- Central Limit Theorem
- Confidence Interval
- Hypothesis Testing
- One Sided – Two Sided Tests
- Chi-Square and Other Test
- Correlation & Introduction to Regression
- Workbook

Data Basics…



Country	2013 GDP	No. of States	Hemisphere	HDI category
Country A	1.83	29	Northern	High
Country B	12.97	51	Southern	Low
Country C	0.87	5	Southern	Low
Country D	2.89	15	Northern	Medium
Country E	1.19	10	Northern	High

2013 GDP in Trillion $ = Continious Numerical
Number of States = Discrete Numerical
Which hemisphere the country lies * = Regular Categorical
Human Development Index Category (Derived from HDI Score) = Ordinal Categorical

Cross Sectional Data

- Its is a collection of observations(behavior) for multiple subjects(entities) at single point in time.
- Max Temperature, Humidity and Wind(all three behaviors) in New York City, SFO, Boston, Chicago(multiple entities) on 1/1/2015(single instance)

Cross Sectional Data

- It is a collection of observations(behavior) for a single subject(entity) at different time intervals(generally equally spaced)
- Max Temperature, Humidity and Wind(all three behaviors) in New York City(single entity) collected on First day of every year(multiple intervals of time)

Panel Data

- It is usually called as Cross-sectional Time-series data as it a combination of above mentioned types, i.e., collection of observations for multiple subjects at multiple instances
- Max Temperature, Humidity and Wind(all three behaviors) in New York City, SFO, Boston, Chicago(multiple entities) on First day of every year(multiple intervals of time)

Population, Sample and Sources of Bias

Population : A population includes each element from the set of observations that can be made.

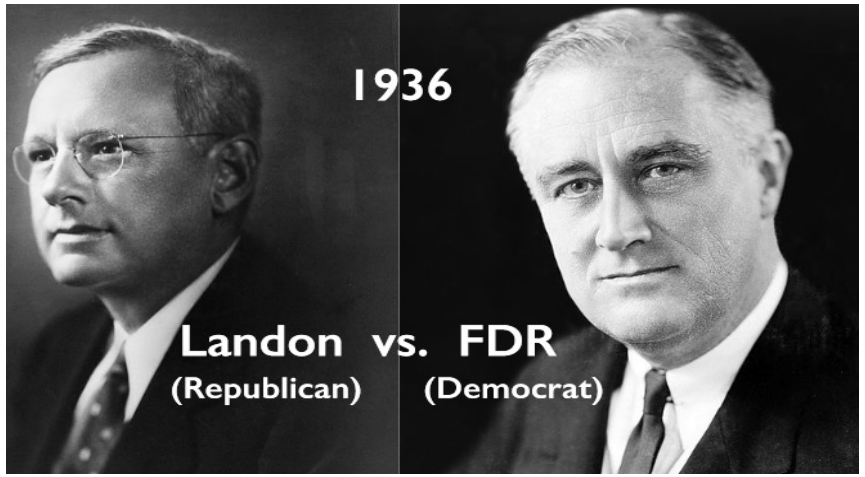
E.g. MNC Employees residing in Bangalore, All Honda cars manufactured in 2014.

Sample : A sample consists only of observations drawn from the population.

E.g. Fidelity Employees from EGL and Manyata, Honda City manufactured in Q2 of 2014.

*In statistics & survey methodology, **sampling** is concerned with the selection of a subset of individuals (sample) from within a statistical population to estimate characteristics of the whole population.*

In cases, when the universe (population) is very large, then the sampling method is the only practical method for collecting the data. It is economical, scientific and consumes less time.



Disadvantages of sampling

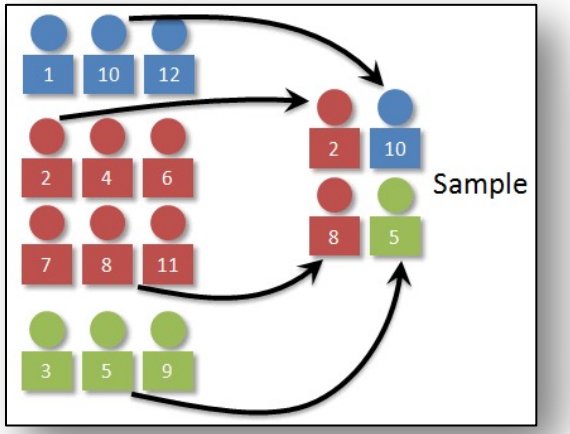
- ❑ Chances of introducing sampling bias.
- ❑ Sampling Error.
- ❑ Will never be an exact representation of the population.

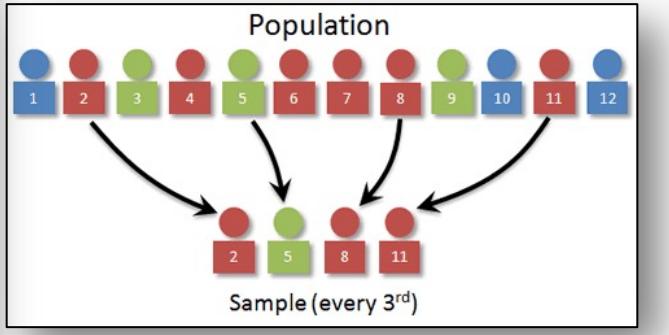
Some Sources of Sampling Bias

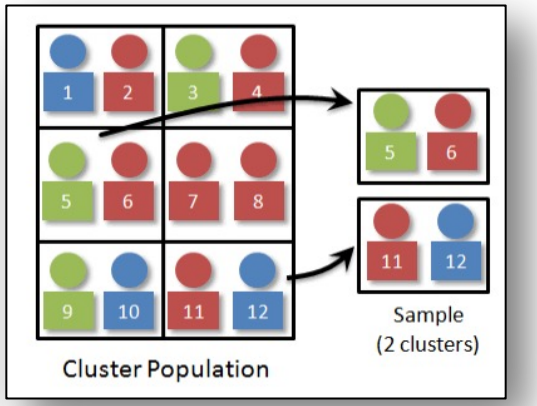
- ❑ **Convenience sample:** Individuals who are easily accessible are more likely to be included in the sample
- ❑ **Non-Response:** If only a (non-random) fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population.
- ❑ **Voluntary Response :** Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue

Popular Sampling Methods

- **Simple random sampling** is the simplest form of sampling. Each member of the population has an equal and known chance of being selected.
- **Systematic sampling** is often used instead of random sampling. It is also called an Nth name selection technique. This involves a random start and then proceeds with the selection of every *k*'th element from then onwards. In this case, *k*=(population size/sample size).
- **Stratified sampling** is commonly used probability method that is superior to random sampling because it reduces sampling error. A stratum is a subset of the population that share at least one common characteristic. Examples of stratums might be males and females, or managers and non-managers. The researcher first identifies the relevant stratums and their actual representation in the population. Random sampling is then used to select a *sufficient* number of subjects from each stratum. "*Sufficient*" refers to a sample size large enough for us to be reasonably confident that the stratum represents the population. Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums.
- **Cluster sampling** is a sampling technique where the entire population is divided into groups, or clusters, and a random sample of these clusters are selected. All observations in the selected clusters are included in the sample.





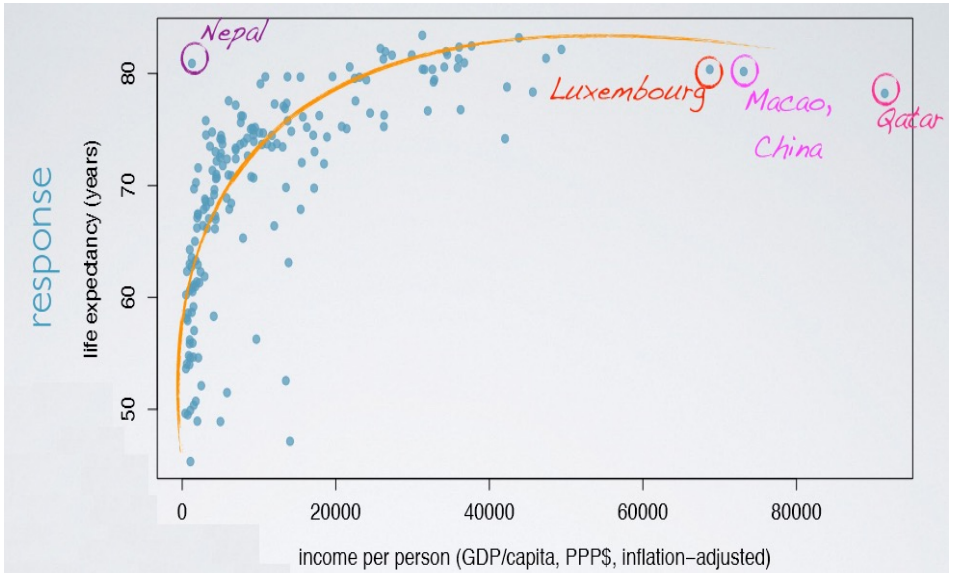


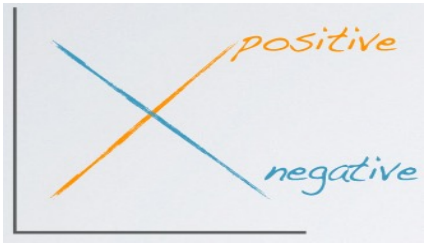
Central Tendency – Measure of Centre

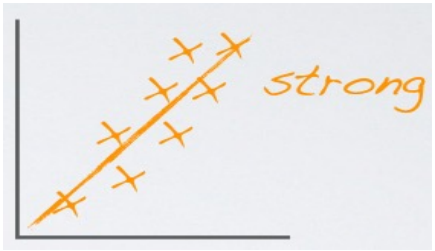
Category	Description	Advantage	Disadvantage	Example
Mean (Arithmetic Mean)	The arithmetic mean of a set of data is found by taking the sum of the data, and then dividing the sum by the total number of values in the set. A mean is commonly referred to as an average.	▪ Simple to Understand/explain ▪ Takes every value into account.	▪ Affected by extreme values	▪ Mean of the series 1-10 would be (1+2+…10)/10 = 5.5
Median	The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value (or highest to lowest) and picking the middle one.	▪ Finds the middle number of a set of data, so outliers have little or no effect	▪ If the gap between some numbers is large, while it is small between other numbers in the data, this can cause the median to be a very inaccurate way to find the middle of a set of values.	▪ Case 1 – Median of the series (1-11) would be - 6 ▪ Case-2 – Median of the series (1-10) would be (5+6)/2 =5.5
Mode	The mode is the value that occurs most frequently in a data set	Allows you to see what value happened the most in a set of data. This can help you to figure out things in a different way. It is also quick and easy.	Could be very far from the actual middle of the data. The least reliable way to find the middle or average of the data.	▪ Example 1 –(1,2,3,4,4,5) - Mode 4 ▪ Example 2 – (1,2,2,3,3,4) – Mode 2,3 ▪ Example 3 – (1,1,2,2,3,3) – No Mode

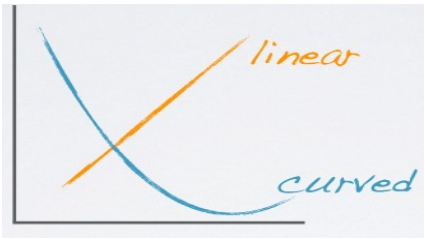
USAGE: Incase the dataset contains outliers, it is suggested that one uses Median instead of Mean as the average reported could be highly skewed by outliers, but again it depends on how we use this data.
E.g. College placement committee would want to lure students by reporting mean salary received by its students but a student seeking admission should consider median as the average reported could be biased by couple of highly compensated international placements.

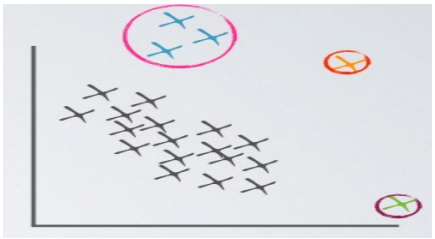
Working on Numerical Data…And Visualizing…!!!

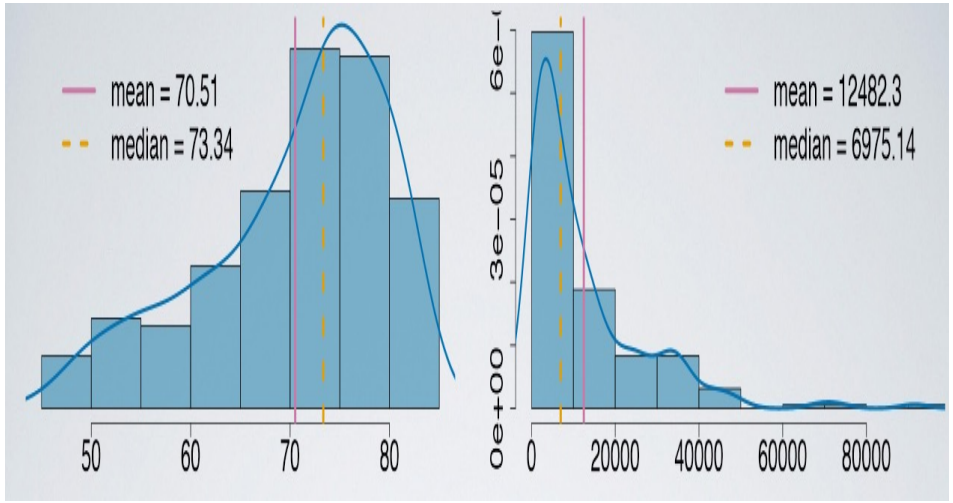


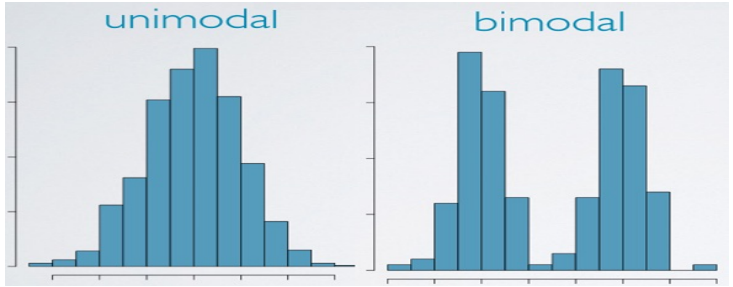


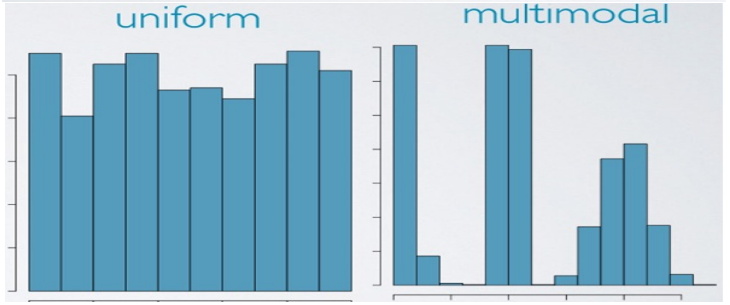












Skewness and Kurtosis!

Skewness: It is the degree of departure from symmetry of a distribution.

•Skewness > 0 - Right skewed distribution - most values are concentrated on left of the mean, with extreme values to the right.

•Skewness < 0 - Left skewed distribution - most values are concentrated on the right of the mean, with extreme values to the left.

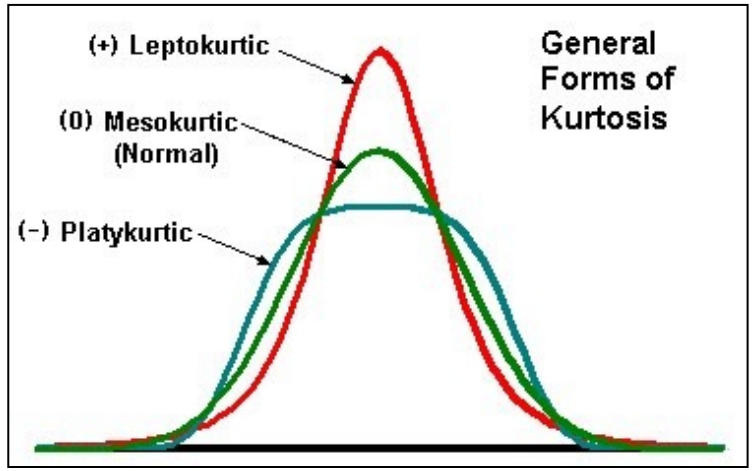
•Skewness = 0 - mean = median, the distribution is symmetrical around the mean.

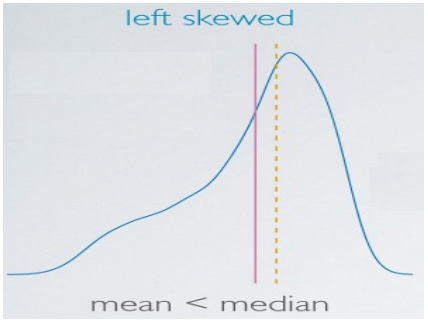
Kurtosis: It is the degree of peakedness of a distribution.

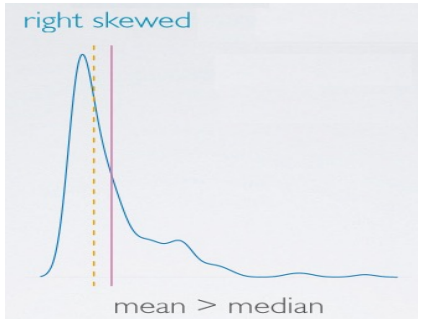
•Kurtosis > 3 - Leptokurtic distribution, sharper than a normal distribution, with values concentrated around the mean and thicker tails. This means high probability for extreme values.

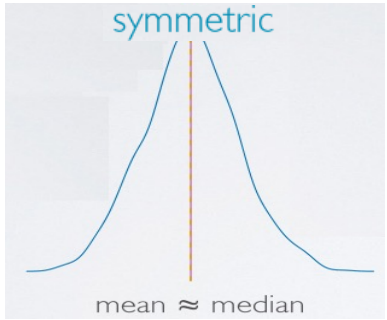
•Kurtosis < 3 - Platykurtic distribution, flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution, and the values are wider spread around the mean.

•Kurtosis = 3 - Mesokurtic distribution - normal distribution for example.





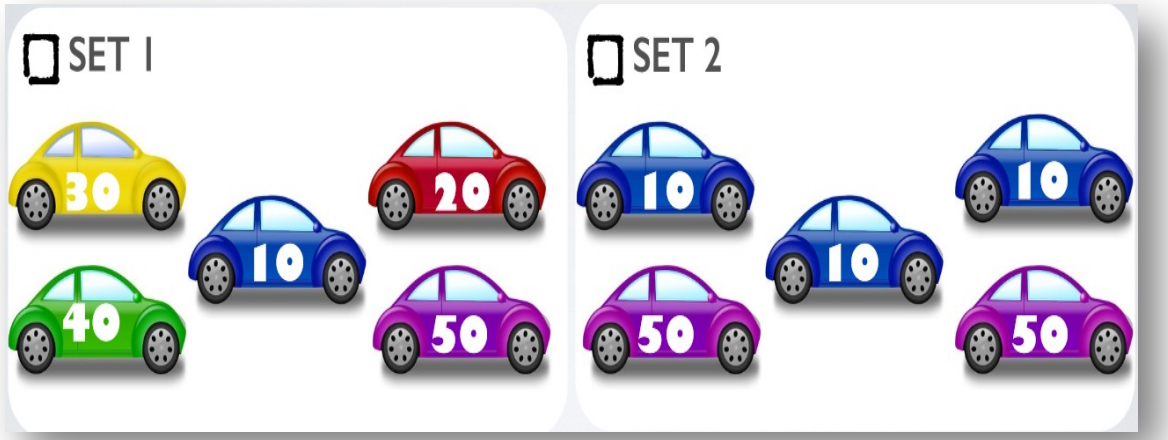




Pattern	Distribution
Mean < Median	Left Skewed
Mean > Median	Right Skewed
Mean = Median	Symmetric

Dispersion – Measure of Spread

- **Range:** The difference between the largest and the smallest value in the dataset. Since the range only takes into account two values from the entire dataset, it may be heavily influenced by outliers in the data.
- **Interquartile range :** Distance between the 25th and 75th percentiles (Q3 - Q1). By definition, this contains 50% of the data points in a dataset.
- **Mean Deviation:** The average of the absolute values of the differences between individual numbers and their mean.
- **Standard deviation (represented by the symbol σ):** shows how much variation or "dispersion" exists from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values. The square of SD is know as **Variance**.
- **Coefficient of Variation:** It's a relative measure. The formula is (SD/Mean)*100



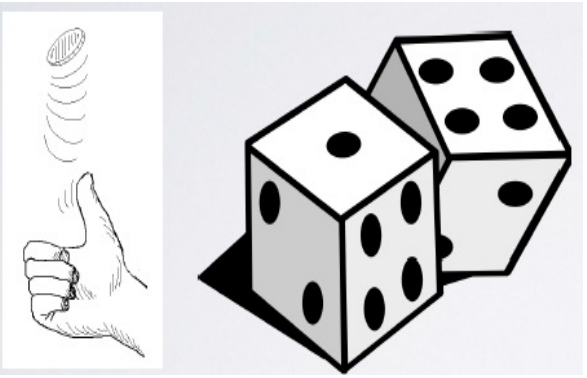
- **Which of the following sets of cars has a more diverse composition of colors?**
- **Which of the following sets of cars has more variable mileage?**

Probability Theory

Probability theory is the branch of mathematics concerned with the analysis of random phenomena/process.

Random Process : Random processes (and variables) talk about quantities and signals which are unknown in advance. In a random process we know what outcomes could happen, but we don't know which particular outcome will happen.

Random Variable, is a variable whose possible values are numerical outcomes of a random phenomenon. There are two types of random variables, discrete and continuous.



- **Probability:** Probability is the measure of the likeliness that an event will occur. Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).
- **Trial:** Any particular performance of a random experiment is called a trial.

  E.g.: Rolling a die, Running a Direct Mail campaign
- **Experiment:** This is considered to be a larger entity formed by the combination of a number of trials.

  E.g.: A game of a 3 coin tosses, Pick a random card from a deck and roll a die.
- **Event/Outcome:** A result that is caused by some previous action. The results or outcomes or observations of an experiment are called events.

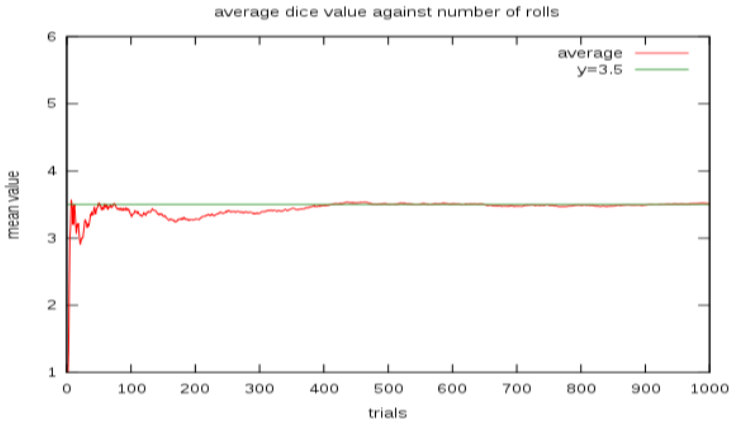
  E.g.: Appearance of six in the roll of a die, # of responders in a campaign

Concepts Related to Probability Theory

- **Law of Large Numbers:**

According to this law, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed.

E.g.: If a large number of six-sided dice are rolled, the average of their values (sometimes called the sample mean) is likely to be close to 3.5, with the precision increasing as more dice are rolled.



- **Sample Space:** This is a collection of all possible outcomes of a trial.
- **Mutually Exclusive/Disjoint Events:** Events are said to be mutually exclusive or disjoint when the occurrence of one of the events is not associated with the occurrence of the other(s).

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B).$$

- **Independent Events:** A set of events are independent if knowing the outcome of one provides no useful information about the outcome of the other.

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B)$$

- **Complementary Events:** Complementary events are two mutually exclusive events whose probabilities add up to 1.

Continued Concepts….Marginal, Joint, Conditional Probabilities

- **Simple or Marginal probability:** The probability of an event occurring (p(A)), which is not conditioned on another event. Example: the probability that a card drawn is red (p(red) = 0.5). Another example: the probability that a card drawn is a 4 = (p(four)=1/13).

- **Joint probability:** p(A and B) The probability of event A and event B occurring. It is the probability of the intersection of two or more events. The probability of the intersection of A and B may be written p(A ∩ B). Example: the probability that a card is a four and red =p(four and red) = 2/52=1/26. (There are two red fours in a deck of 52, the 4 of hearts and the 4 of diamonds).

- **Conditional probability:** p(A|B) is the probability of event A occurring, given that event B occurs. Example: given that you drew a red card, what's the probability that it's a four (p(four|red))=2/26=1/13. So out of the 26 red cards (given a red card), there are two fours so 2/26=1/13.

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

The conditional probability of A given B is equal to the joint probability of A and B divided by the marginal of B.

- **Bayes' theorem:** An equation that allows us to manipulate conditional probabilities. For two events, A and B, Bayes' theorem lets us to go from p(B|A) to p(A|B) if we know the marginal probabilities of the outcomes of A and the probability of B, given the outcomes of A.

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

Bayes Theorem: An example…

In Boston, 51% of the adults are males. One adult is randomly selected for a survey involving credit card usage.

- a) Find the probability that the selected person is a male.
- b) It is later learned that the selected survey subject was smoking a cigar. Also, 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars. Use this additional information to find the probability that the selected subject is a male.

Solution:

Notations:

M = male	$\overline{M}$ = female (or not male)
C = cigar smoker	$\overline{C}$ = not a cigar smoker.

- a) Before using the information given in part b, we know only that 51% of the adults in Boston are males, so the probability of randomly selecting an adult and getting a male is given by P(M) = 0.51.
- b) Using additional information provided, we get:

$P(M) = 0.51$	because 51% of the adults are males
$P(\overline{M}) = 0.49$	because 49% of the adults are females (not males)
$P(C\|M) = 0.095$	because 9.5% of the males smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a male, is 0.095.)
$P(C\|\overline{M}) = 0.017.$	because 1.7% of the females smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a female, is 0.017.)

Using Bayes Theorem we have:

$$P(M \mid C) = \frac{P(M)\cdot P(C\mid M)}{[P(M)\cdot P(C\mid M)]+[P(\overline{M})\cdot P(C\mid \overline{M})]}$$

$$= \frac{0.51\cdot 0.095}{[0.51\cdot 0.095]+[0.49\cdot 0.017]}$$

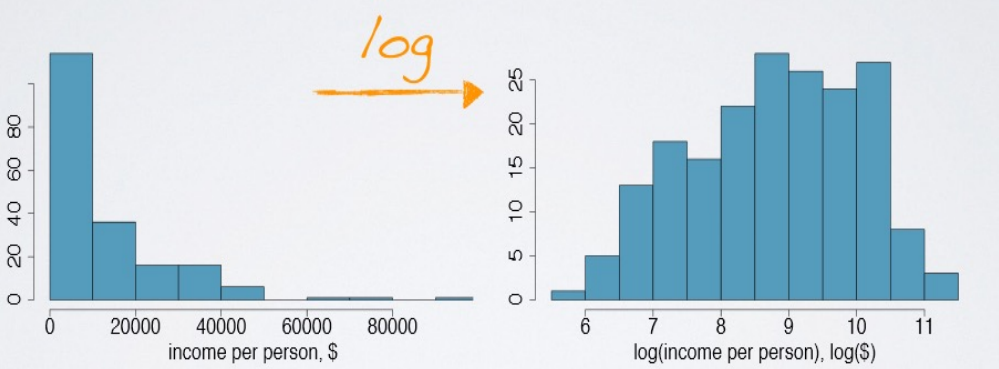
$$= 0.85329341$$

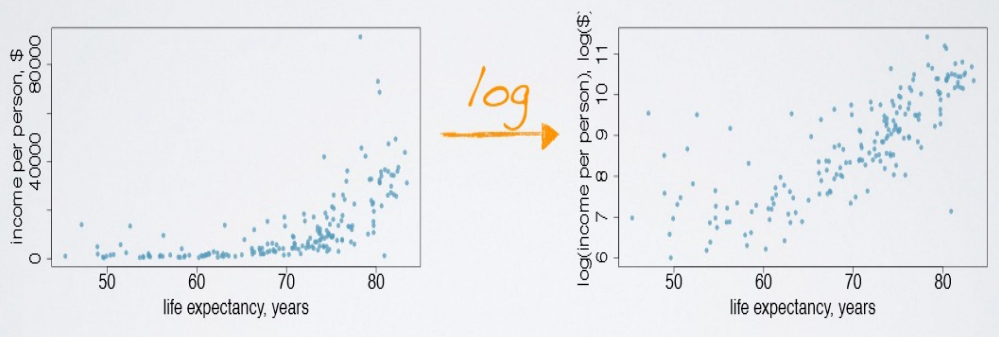
$$= 0.853 \text{ (rounded)}$$

Before we knew that the survey subject smoked a cigar, there is a 0.51 probability that the survey subject is male. However, after learning that the subject smoked a cigar, we revised the probability to 0.853. There is a 0.853 probability that the cigar-smoking respondent is a male. This makes sense, because the likelihood of a male increases dramatically with the additional information that the subject smokes cigars (because many more males smoke cigars than females).

Variable Transformation

- A transformation is a rescaling of the data using a function
- When data are very strongly skewed, we sometimes transform them so they are easier to model
- Transformations can also be applied to one or both variables in a scatterplot to make the relationship between the variables more linear, and hence easier to model with simple methods





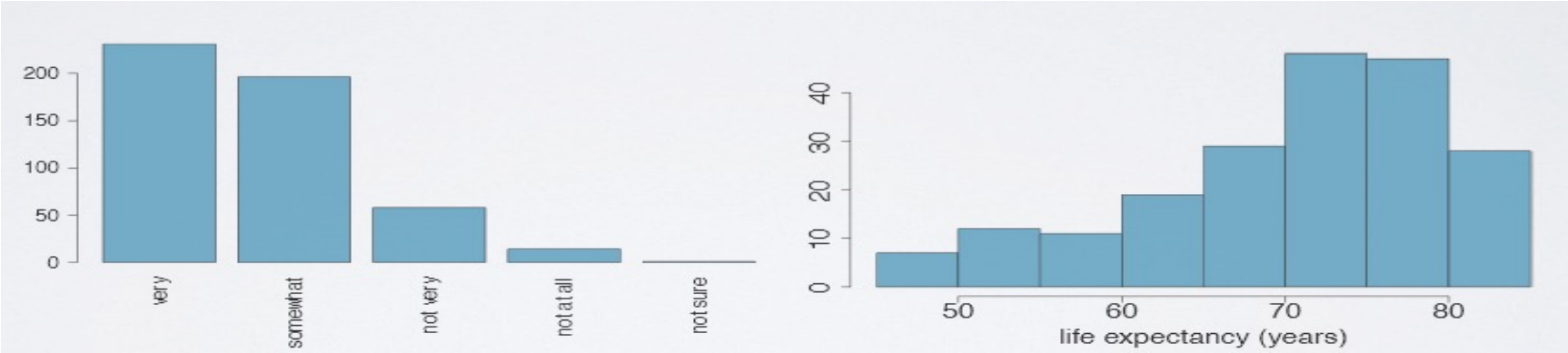
Variable Transformation

Time_Period	World Auto Sales	Log_World_Auto_Sales	ln_World_Auto_Sales	Sq_World_Auto_Sales
January2000	2.578355	0.947151599	0.387844187	6.647914506
February2000	2.847629	1.046486718	0.351169341	8.108990922
March2000	3.395962	1.222587079	0.29446737	11.53255791
April2000	2.870469	1.054475431	0.348375126	8.23959228
May2000	3.171373	1.154164617	0.315320841	10.05760671
June2000	3.05904	1.118101141	0.326899942	9.357725722
July2000	2.744722	1.009679795	0.364335623	7.533498857
August2000	2.858729	1.05037712	0.349805805	8.172331495
September2000	3.032695	1.109451663	0.32973972	9.197238963
October2000	2.647959	0.973789155	0.377649352	7.011686866
November2000	2.61578	0.961562332	0.382295147	6.842305008
December-00	2.484422	0.910040037	0.402508109	6.172352674

- We generally use 6 types of transformations – log, inverse, cube, cube-root, square, square-root to generate better explanatory power
- Log, square-root & inverse transform positively skewed data into a more symmetric form.

Exploring Categorical Variables

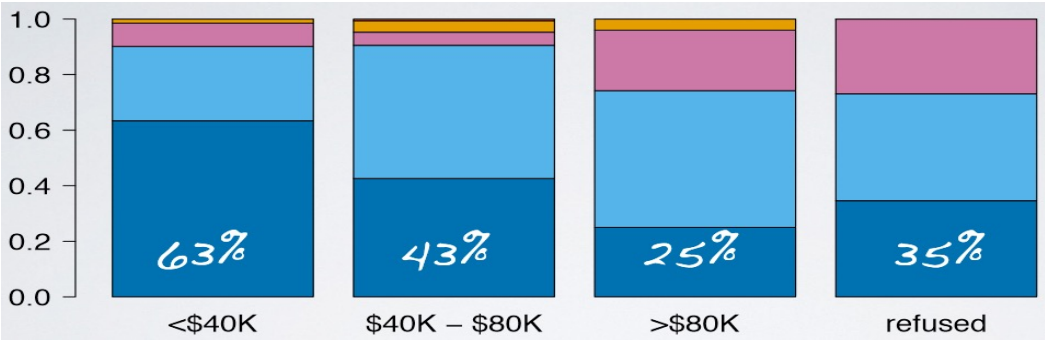
- A 2014 poll in the US asked respondents how difficult they think it is to save money



	Counts	Frequencies
Very	231	46%
Somewhat	196	39%
Not very	58	12%
Not at all	14	3%
Not sure	1	~0%
Total	500	100%

		Income				
		< $40K	$40K - $80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

		Income				
		< $40K	$40K - $80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

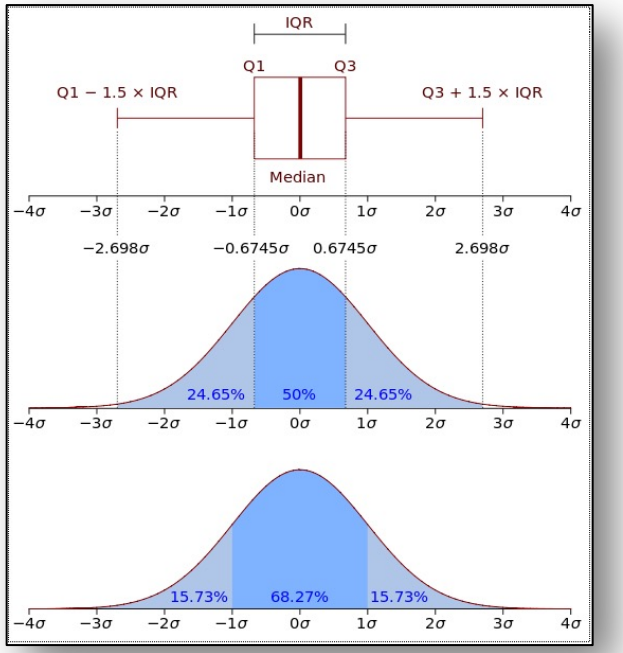


- Useful for visualizing conditional frequency distributions
- In order to explore the relationship between the variables, we need to compare relative frequencies

Distribution…parameter…statistics !!

A **Probability Distribution** assigns a probability to each measurable subset of the possible outcomes of a random experiment, survey, or procedure of statistical inference.
In probability theory, a **Probability Density Function** (PDF), or density of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value. The probability of the random variable falling within a particular range of values is given by the integral of this variable's density over that range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to one.

Continuous distributions describe an infinite number of possible data values.
Example: Normal Distribution, Chi-Square Distribution, Student's T Distribution
Discrete distributions describe a finite number of possible values.
Example: Binomial Distribution, Poisson Distribution



Parameter	Statistic
Describes a characteristic of a population	Describes a characteristic of a sample
Fixed Value within the population	Value varies from sample to sample within the same population
Generally represented as Greek alphabets (e.g. Σ)	Generally represented as English alphabets (e.g. S)

Example:

A properly chosen sample of 1600 people across the country was asked if they regularly watch a certain television program, and 24% said yes.

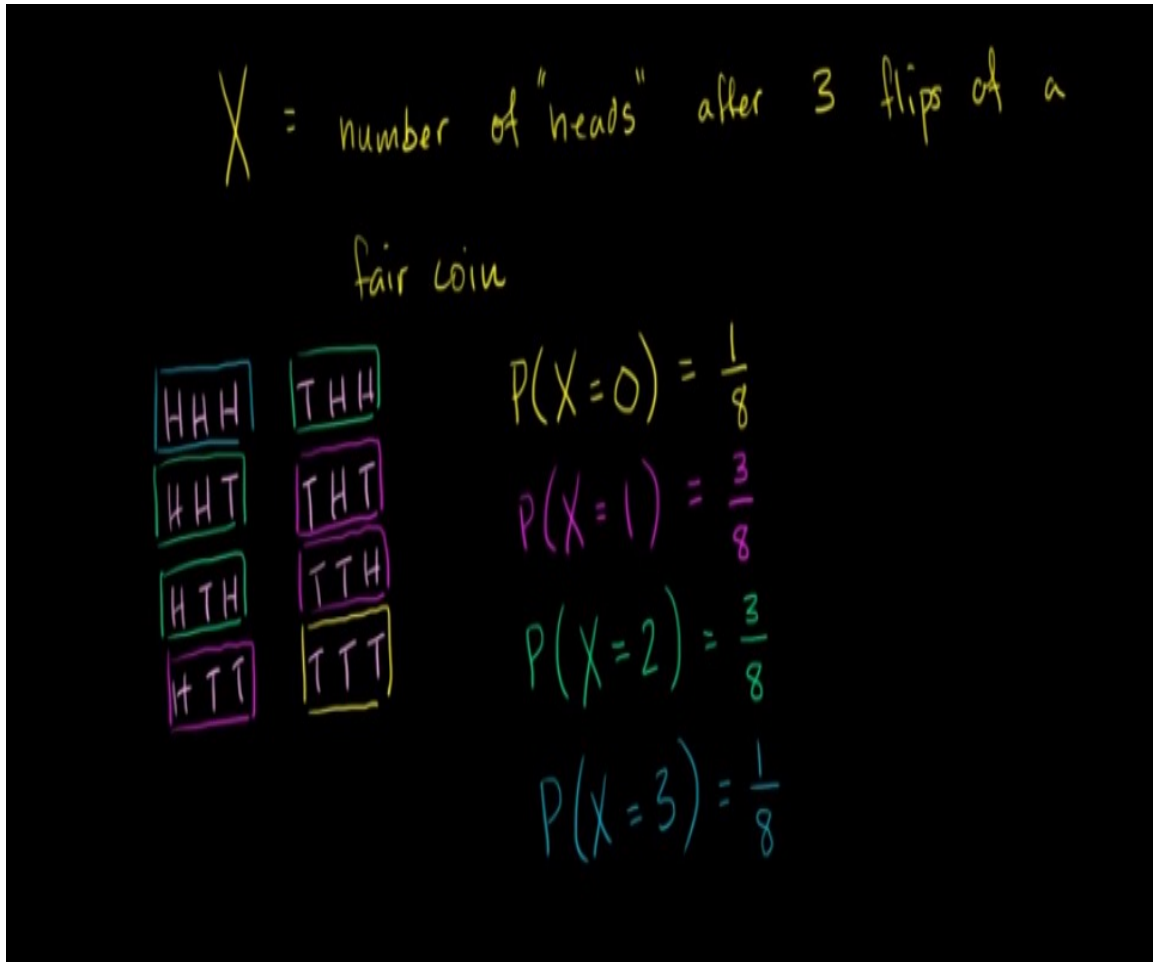
Parameter : The true proportion of all people in the country who watch the program

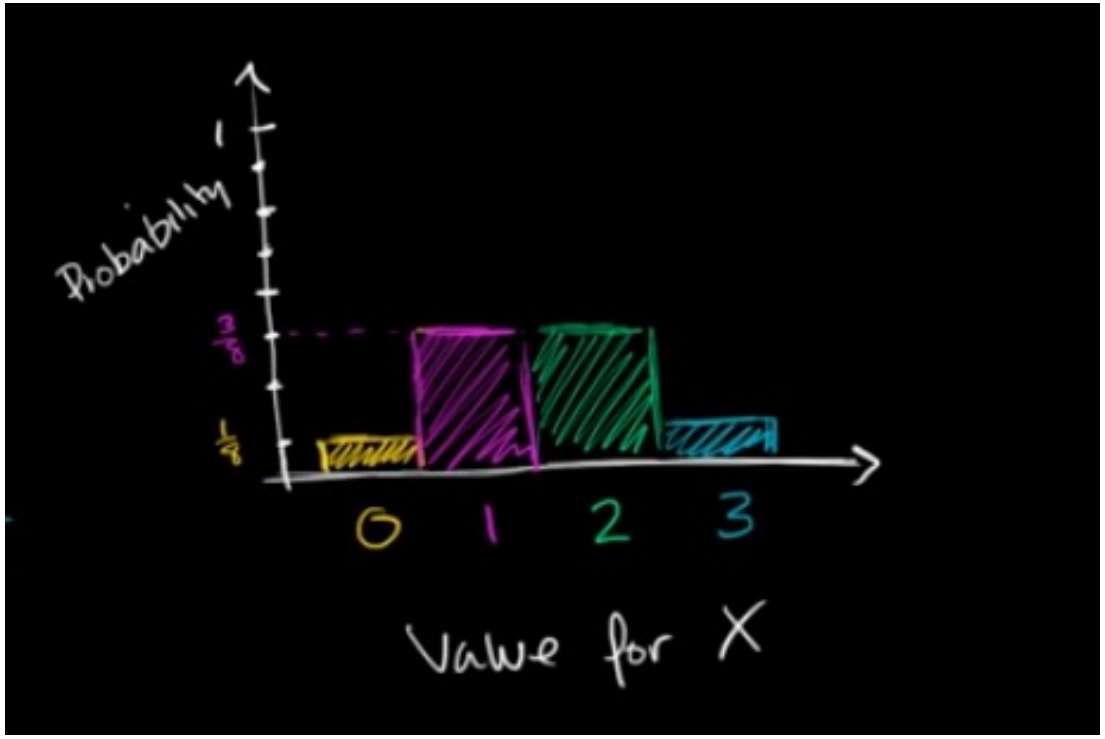
Statistic : 24% (as obtained from the sample of 1600 people)

- **Parameters** are numbers that summarize data for an entire population.
- **Statistics** are numbers that summarize data from a sample, i.e. some subset of the entire population.

Parameters are usually unknown and estimated using statistics from sample.

Discrete Probability Distribution: Probability Mass Function





Discrete Probability Distribution: Example

) plans to buy packs of baseball cards until he gets the card of his favorite player, but he only enough money to buy at most 4 packs. Suppose that each pack has probability 0.2 of containing the card Hugo is hoping for.

Let the random variable X be the number of packs of cards Hugo buys. Here is the probability distribution for X:

$X = \#$ of packs	1	2	3	4
$P(X)$	0.2	0.16	0.128	?

Find the indicated probability.

$P(X \geq 2) =$ ☐

Expectations: Mean & Variance (1/2)

12.3 Expectation—Mean, Variance, Moments (discrete distribution)

Let a discrete random variable x assume the values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ respectively. Then the *expectation*, or *'expected value'* of x—written $E(x)$—is defined as the sum of products of the different values of x and the corresponding probabilities.

$$E(x) = \Sigma p_i x_i \qquad (12.3.1)$$

The expected value of x^2 is similarly defined as the sum of products of the squares of values and the corresponding probabilities.

$$E(x^2) = \Sigma p_i x_i^2 \qquad (12.3.2)$$

In general, the expected value of any function $g(x)$ is defined as

$$E[g(x)] = \Sigma p_i g(x_i)$$

Hence, the expected value of a constant k is the constant k itself.

$$E(k) = k,$$

where k is a constant; because $E(k) = \Sigma p_i k = k\Sigma p_i = k$.

If the p.m.f. $f(x)$ is given, then the expectations are defined as follows:

$$E(x) = \Sigma x.f(x) \qquad (12.3.3)$$

$$E(x^2) = \Sigma x^2.f(x) \qquad (12.3.4)$$

Mean of a probability distribution is the expected value of x.

$$\text{Mean } (\mu) = E(x) \qquad (12.3.5)$$

Variance is the expected value of $(x-\mu)^2$, where μ is the mean.

$$\text{Variance } (\sigma^2) = E(x-\mu)^2$$

It may be shown that

$$\sigma^2 = E(x^2) - \mu^2 \qquad (12.3.6)$$

Expectations: Mean & Variance (1/2)

Example 12 : 2 A random variable has the following probability distribution :

x	4	5	6	8
Probability	0.1	0.3	0·4	0.2

Find the expectation and the standard deviation of the random variable. [*C.U., B. Com. (Hons) '80*]

Solution : (Note that the sum of the probabilities is 1 ; $\Sigma p_i = 1$).

$$\text{Mean } (\mu) = E(x) = \Sigma p_i x_i$$
$$= (0.1\times4) + (0.3\times5) + (0.4\times6) + (0.2\times8)$$
$$= 0.4 + 1.5 + 2.4 + 1.6 = 5.9$$

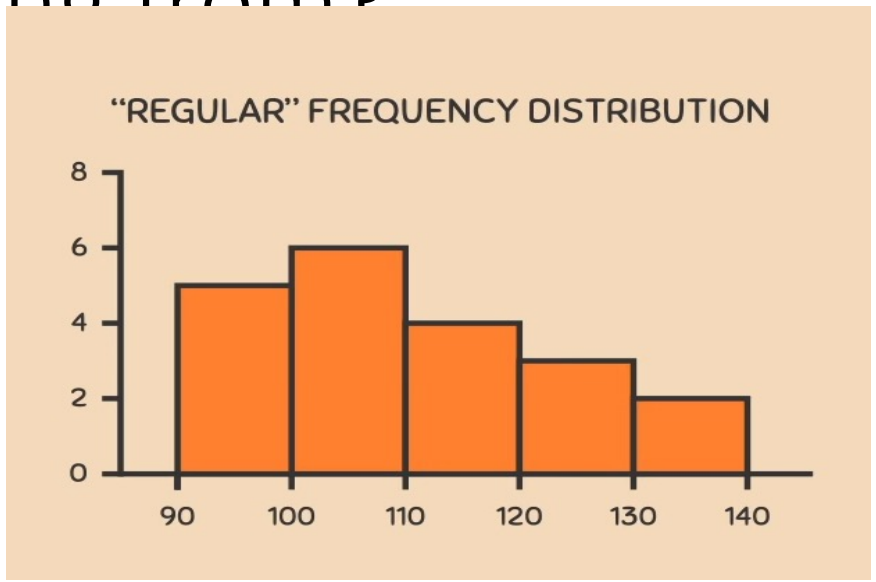
$$E(x^2) = \Sigma p_i x_i^2 = (0.1\times4^2) + (0.3\times5^2) + (0.4\times6^2) + (0.2\times8^2)$$
$$= 1.6 + 7.5 + 14\,4 + 12.8 = 36.3$$

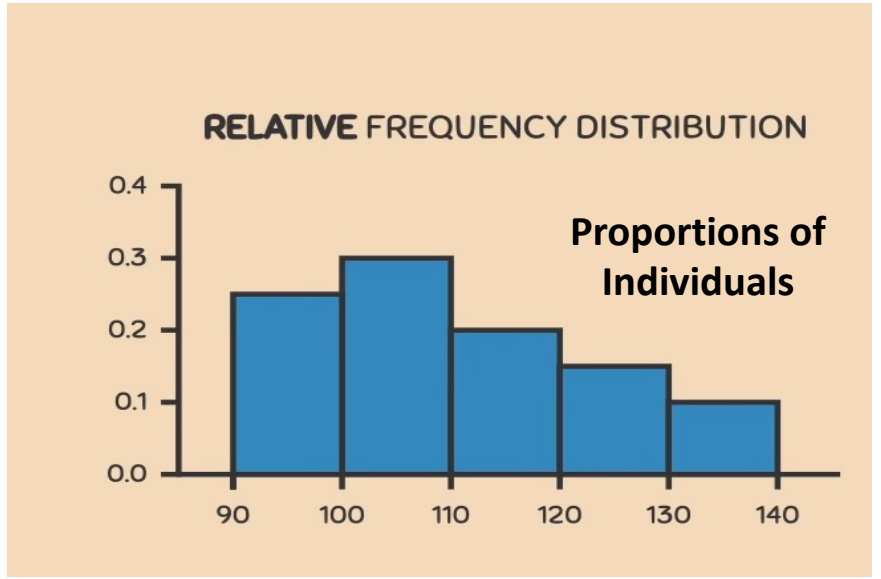
$$\text{Variance } (\sigma^2) = E(x^2) - \mu^2, \text{ using } (12.3.6)$$
$$= 36.3 - (5.9)^2 = 36.3 - 34.81 = 1.49$$

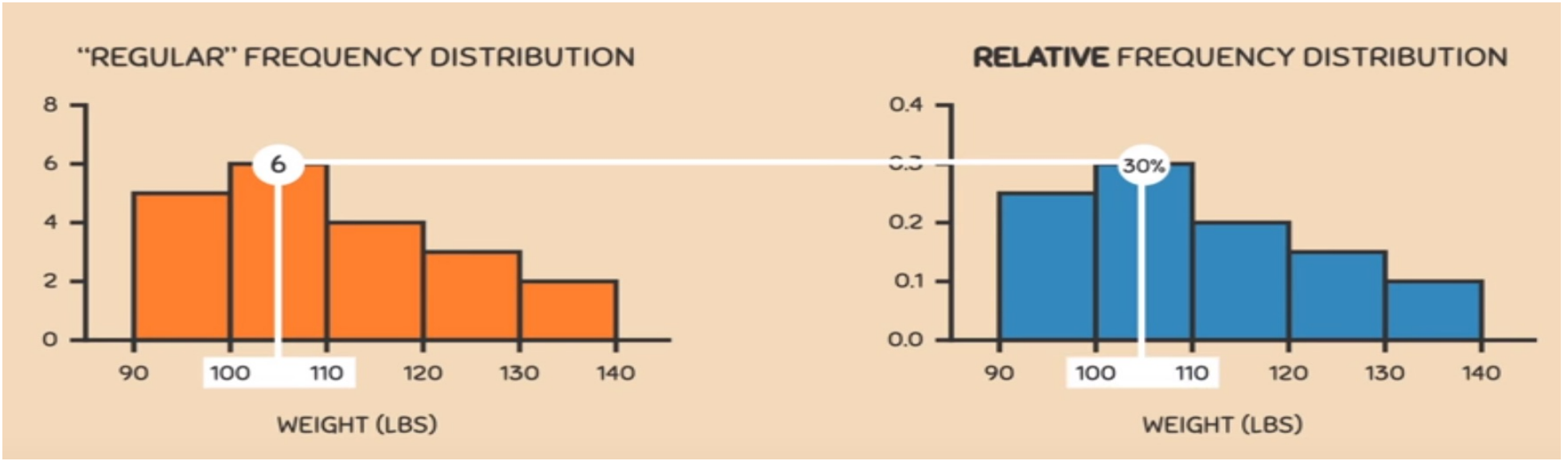
$$\text{S. D. } (\sigma) = \sqrt{1.49} = 1.22$$

Density Curves – What are they & Where they are coming from?

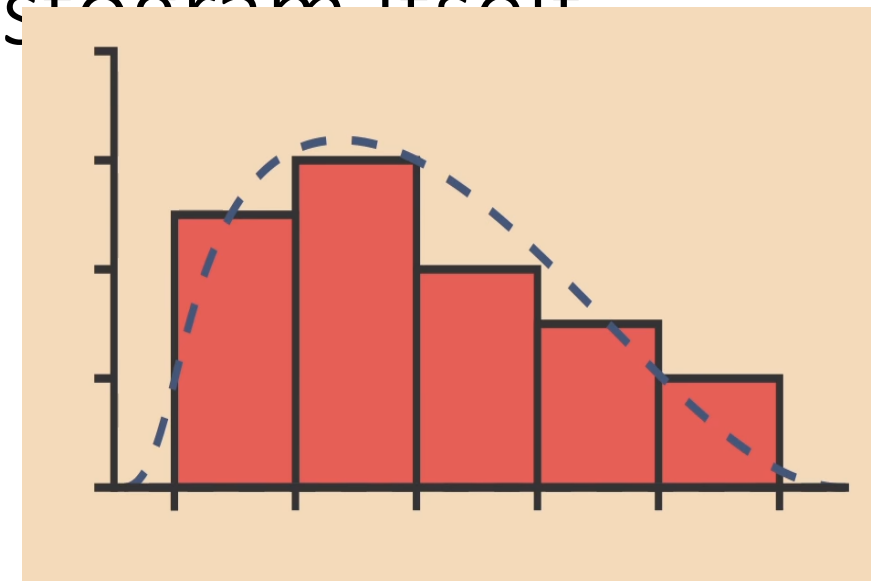
N = 20
Weights in lbs

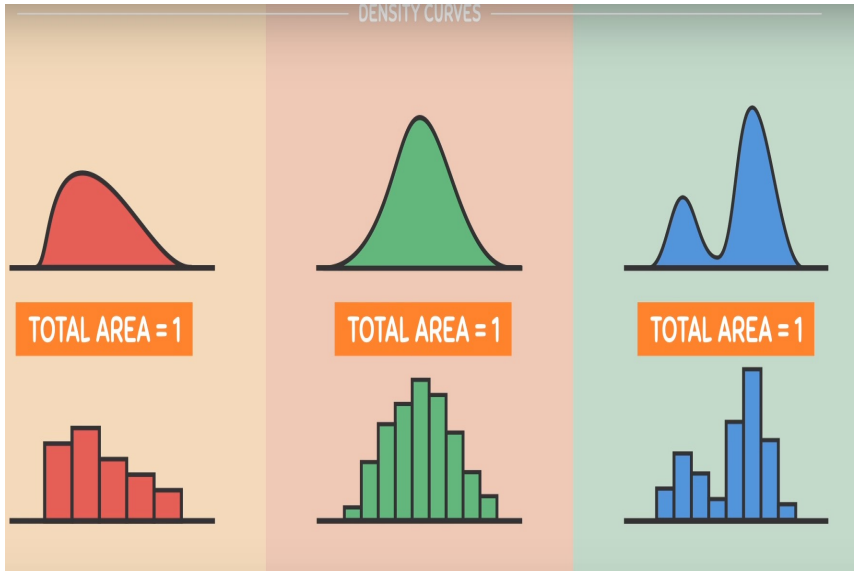


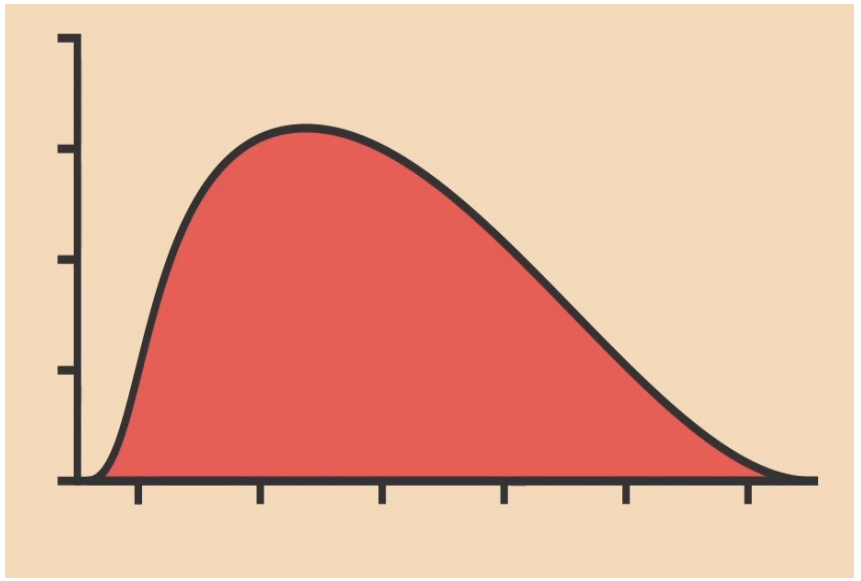




Where it is coming from – A Curve around the histogram itself



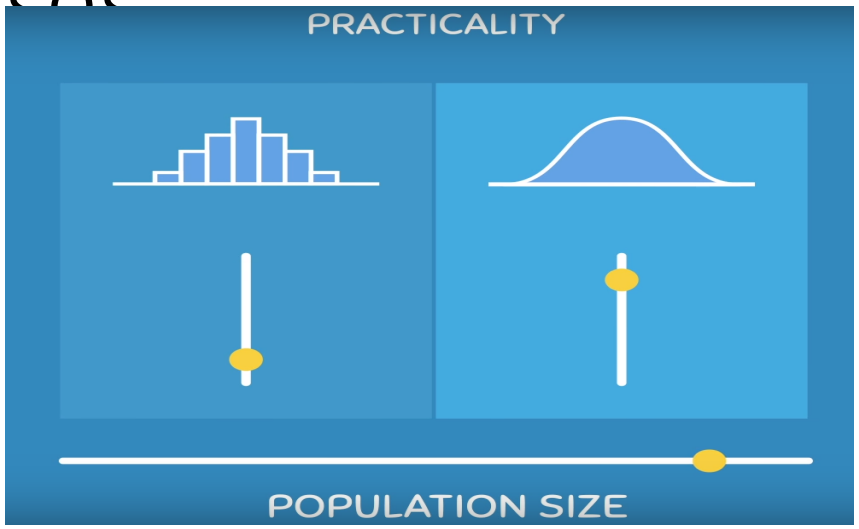


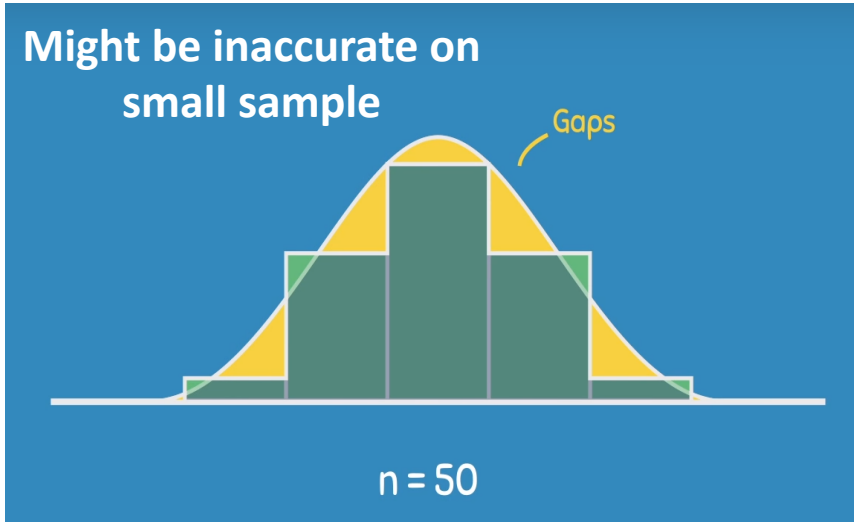


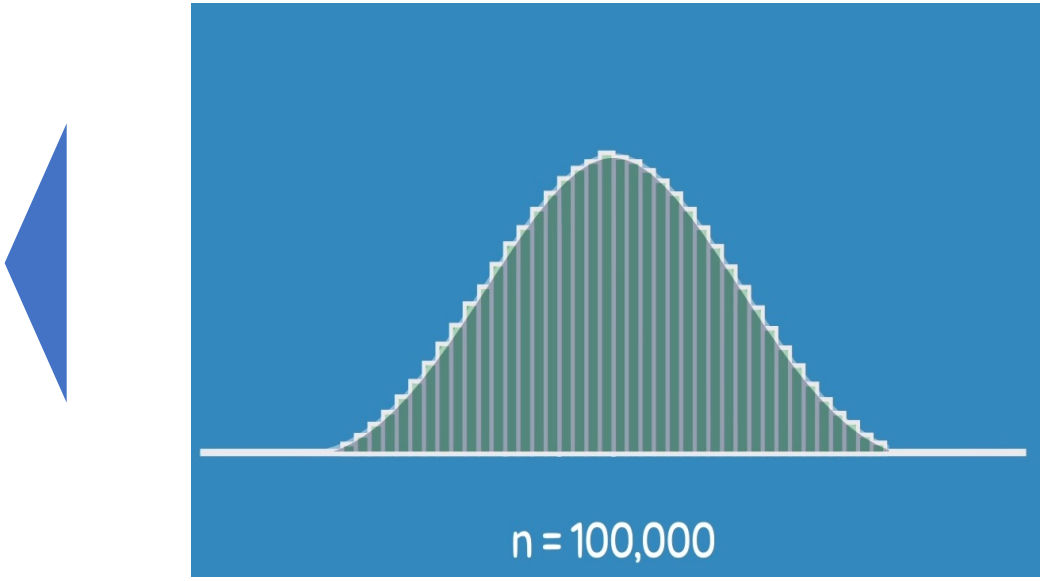
Uses of Histograms

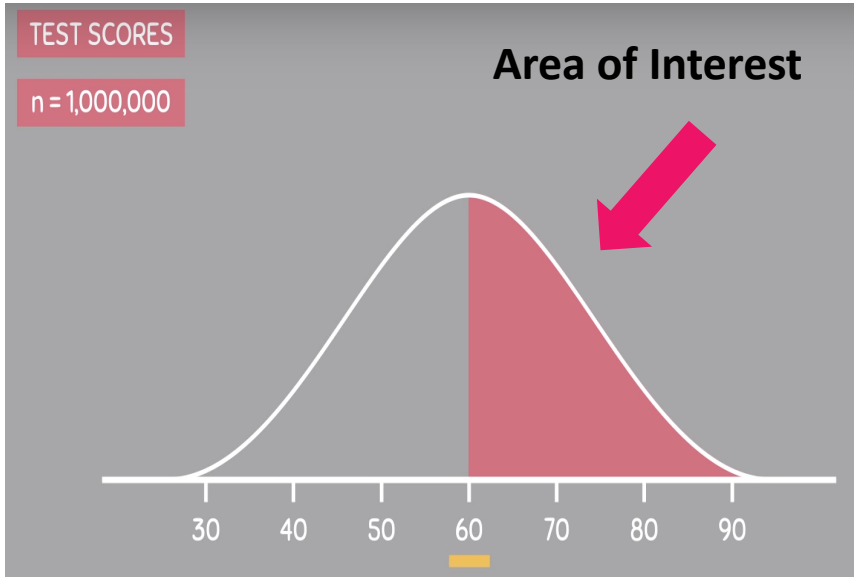
1) **Histograms depends upon how we construct it and depends upon number of intervals we are considering**
2) **Density curves consider infinite amount of intervals – Independent**
3) **Uses of density curves – more popular**

Uses of Density Curves – As Population/Sample size increases

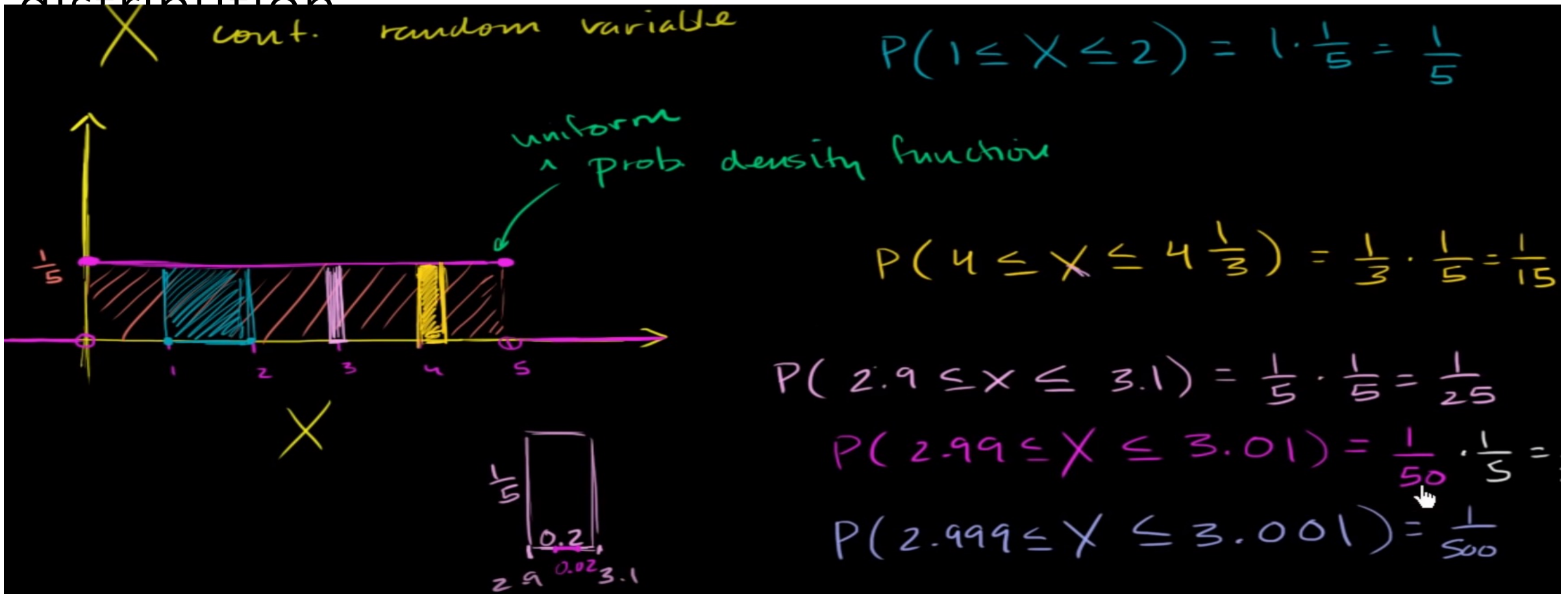


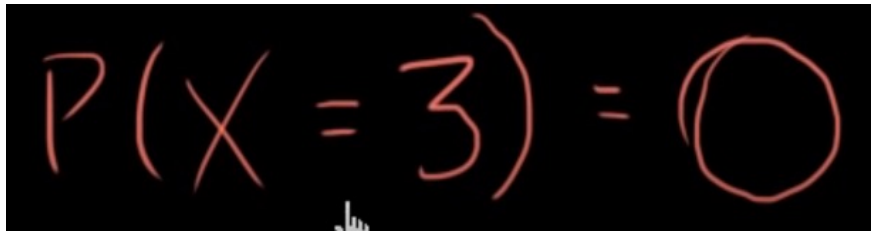






Continuous Probability Distribution Contd. – Area of a distribution





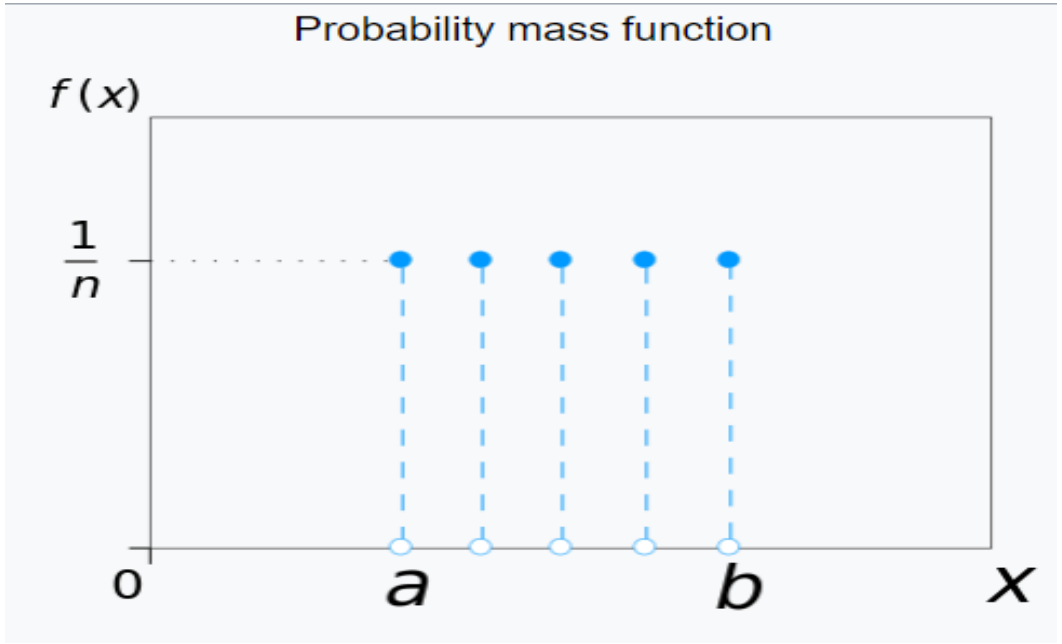
Area = 1 or 100%

Probability of occurrence of any exact value = 0

Uniform Distribution Discrete

In probability theory and statistics, the discrete uniform distribution is a symmetric probability distribution whereby a finite number of values are equally likely to be observed; every one of n values has equal probability 1/n. Another way of saying "discrete uniform distribution" would be "a known, finite number of outcomes equally likely to happen".

A simple example of the discrete uniform distribution is throwing a fair dice. The possible values are 1, 2, 3, 4, 5, 6, and each time the dice is thrown the probability of a given score is 1/6. If two dice are thrown and their values added, the resulting distribution is no longer uniform since not all sums have equal probability. Although it is convenient to describe discrete uniform distributions over integers, such as this, one can also consider discrete uniform distributions over any finite set. For instance, a random permutation is a permutation generated uniformly from the permutations of a given length, and a uniform spanning tree is a spanning tree generated uniformly from the spanning trees of a given graph.



Mean	$\frac{a+b}{2}$
Median	$\frac{a+b}{2}$
Mode	N/A
Variance	$\frac{(b-a+1)^2-1}{12}$

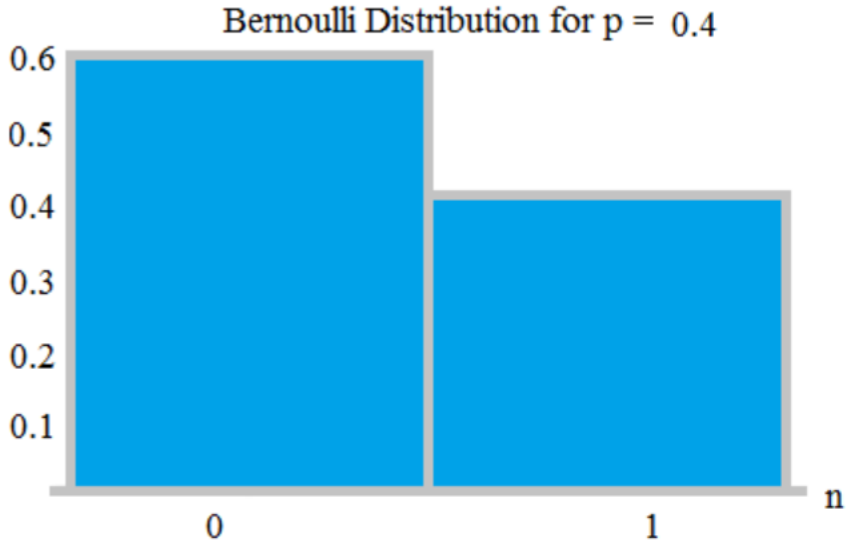
Bernoulli Distribution

A Bernouilli distribution is a discrete probability distribution for a Bernouilli trial — a random experiment that has only two outcomes (usually called a “Success” or a “Failure”)

For example, the probability of getting a heads (a “success”) while flipping a coin is 0.5. The probability of “failure” is 1 – P (1 minus the probability of success, which also equals 0.5 for a coin toss).

It is a special case of the binomial distribution for n = 1. In other words, it is a binomial distribution with a single trial (e.g. a single coin toss).

The probability of a failure is labeled on the x-axis as 0 and success is labeled as 1. In the following Bernoulli distribution, the probability of success (1) is 0.4, and the probability of failure (0) is 0.6:



The expected value for a random variable, X, from a Bernoulli distribution is:

$E[X] = p$.

For example, if $p = .04$, then $E[X] = 0.4$.

The variance of a Bernoulli random variable is:

$Var[X] = p(1 - p)$.

Binomial Distribution

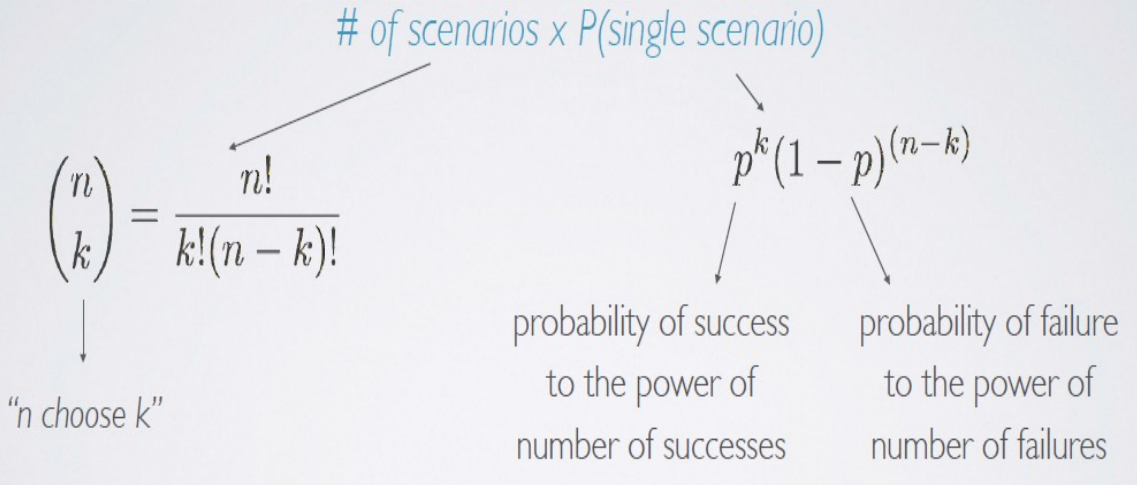
Binomial Distribution: This probability distribution describes the probability of having exactly ‘k’ successes in ‘n’ independent Bernoulli trials with probability of success ‘p’

Binomial distribution:

If p represents probability of success, $(1-p)$ represents probability of failure, n represents number of independent trials, and k represents number of successes

$$\text{P(k successes in n trials)} = \binom{n}{k} p^k (1-p)^{(n-k)}$$

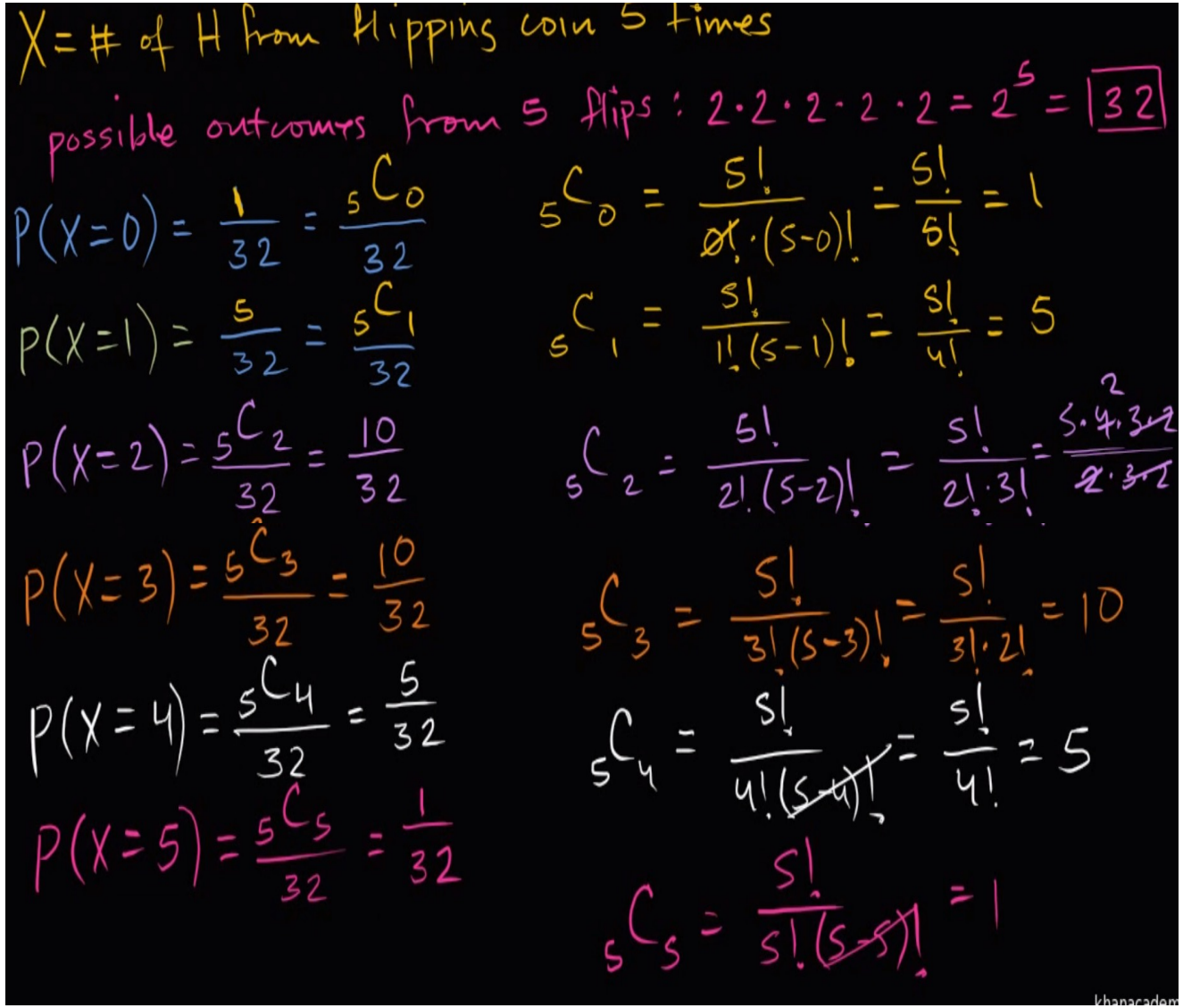
where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

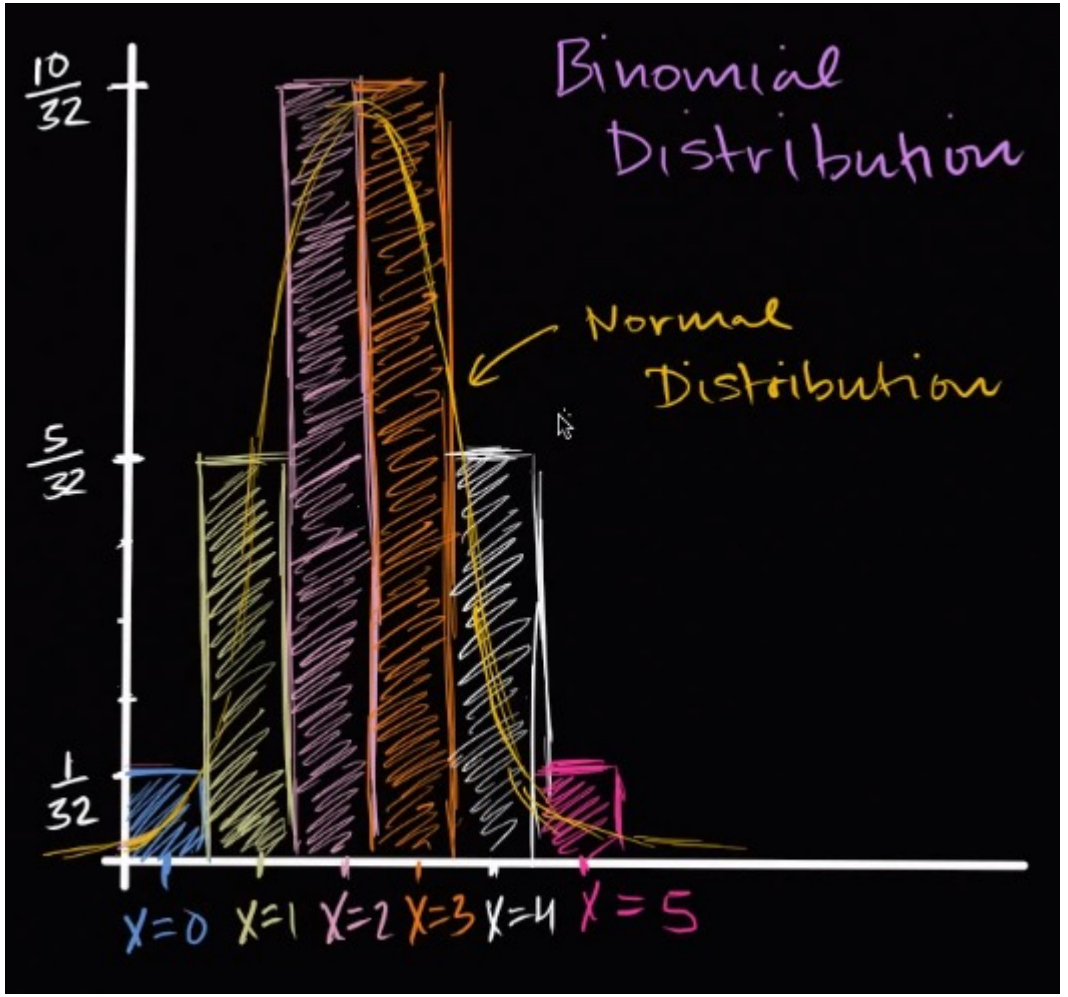


The binomial distribution has the following properties:

- The mean of the distribution (μ_x) is equal to $n * P$.
- The variance (σ^2_x) is $n * P * (1 - P)$.
- The standard deviation (σ_x) is sqrt[$n * P * (1 - P)$].

Binomial Distribution - Example



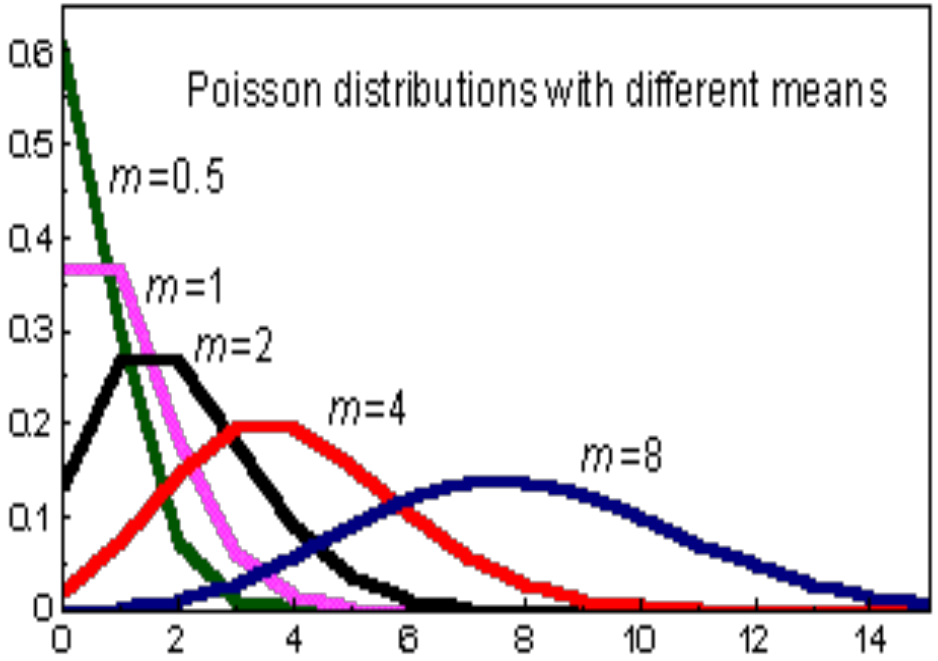


Poisson Distribution

Poisson Distribution: A probability distribution that arises when counting the number of occurrences of a rare event in a long series of trials.

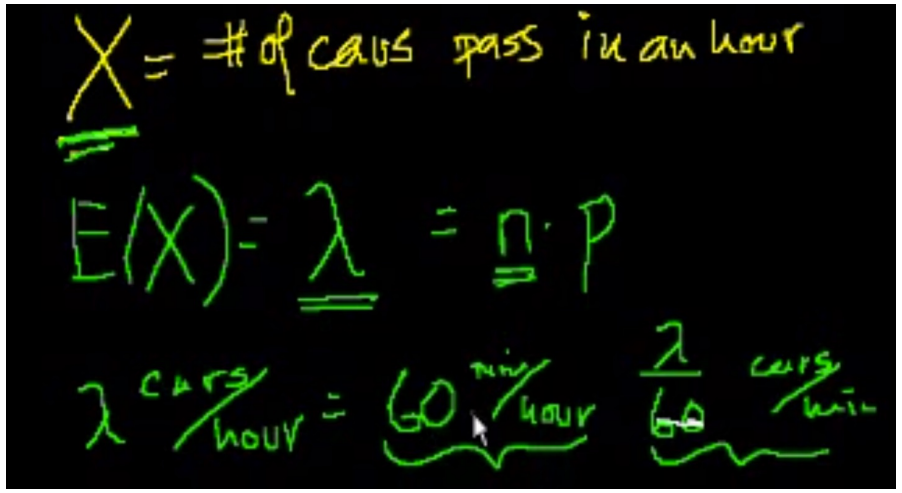
$$P(X = n) = \frac{m^n e^{-m}}{n!}$$

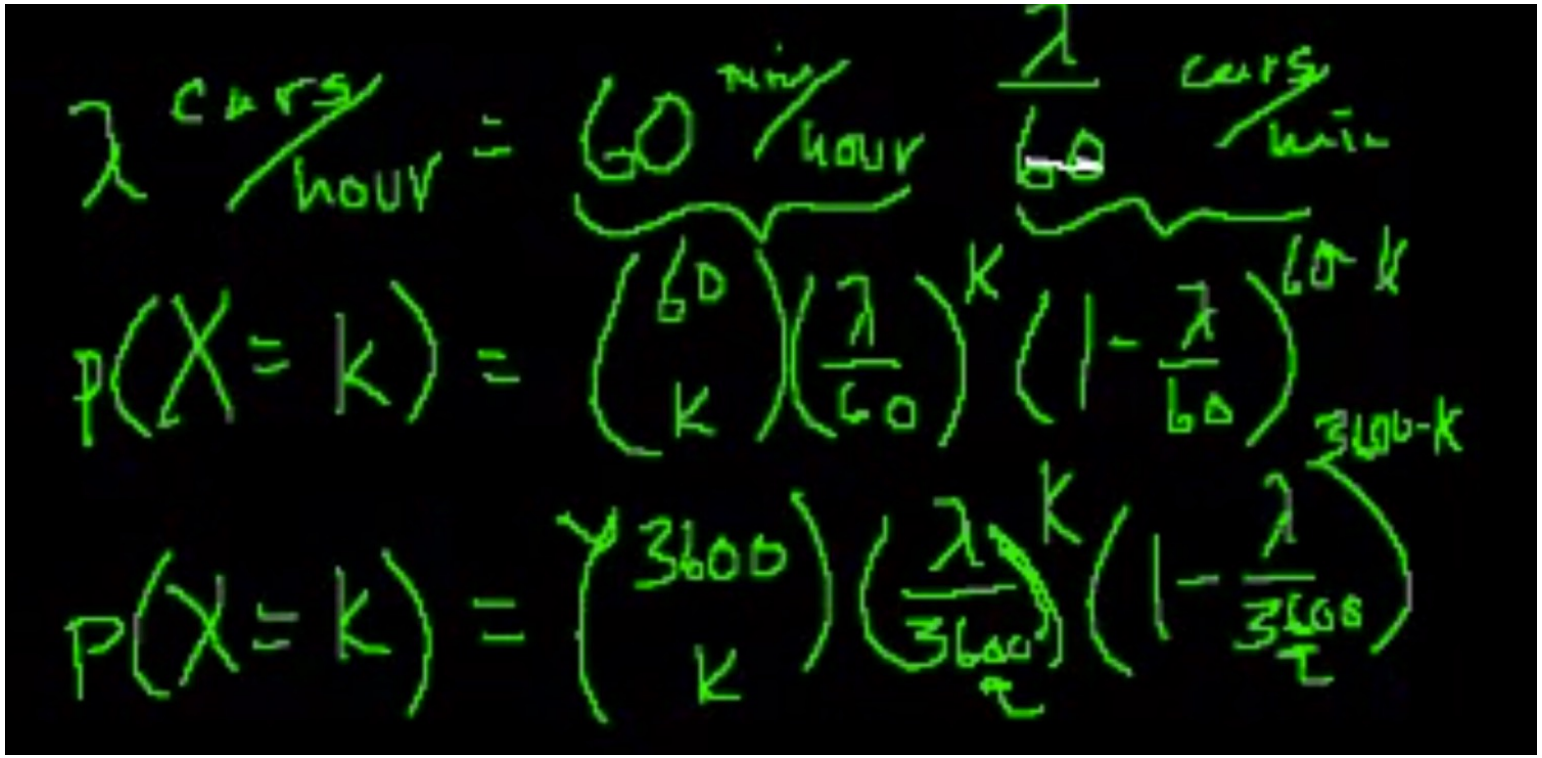
Where --- m : average rate of value
x : Poisson random variable
e : Base of logarithm(e=2.718)



- Used to model the rates of occurrence of an event , that is the number of occurrences in a unit of measure
- Events can happen at any point along a continuum
- At any particular point, the probability of an event is small
- The average number of events is constant over a unit of measure
- Independent events
- One parameter (m : the average number of events in a unit of measure)

Poisson Distribution – Example (1/2)





Poisson Distribution – Example (2/2)

A **Poisson random variable** is the number of successes that result from a Poisson experiment. The probability distribution of a Poisson random variable is called a **Poisson distribution**.

Given the mean number of successes (μ) that occur in a specified region, we can compute the Poisson probability based on the following formula:

Poisson Formula. Suppose we conduct a Poisson experiment, in which the average number of successes within a given region is μ. Then, the Poisson probability is:

$$P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$$

where x is the actual number of successes that result from the experiment, and e is approximately equal to 2.71828.

The Poisson distribution has the following properties:

- The mean of the distribution is equal to μ .
- The variance is also equal to μ .

Poisson Distribution Example

The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Solution: This is a Poisson experiment in which we know the following:

- μ = 2; since 2 homes are sold per day, on average.
- x = 3; since we want to find the likelihood that 3 homes will be sold tomorrow.
- e = 2.71828; since e is a constant equal to approximately 2.71828.

We plug these values into the Poisson formula as follows:

$$P(x; \mu) = (e^{-\mu}) (\mu^{x}) / x!$$

$$P(3; 2) = (2.71828^{-2}) (2^{3}) / 3!$$

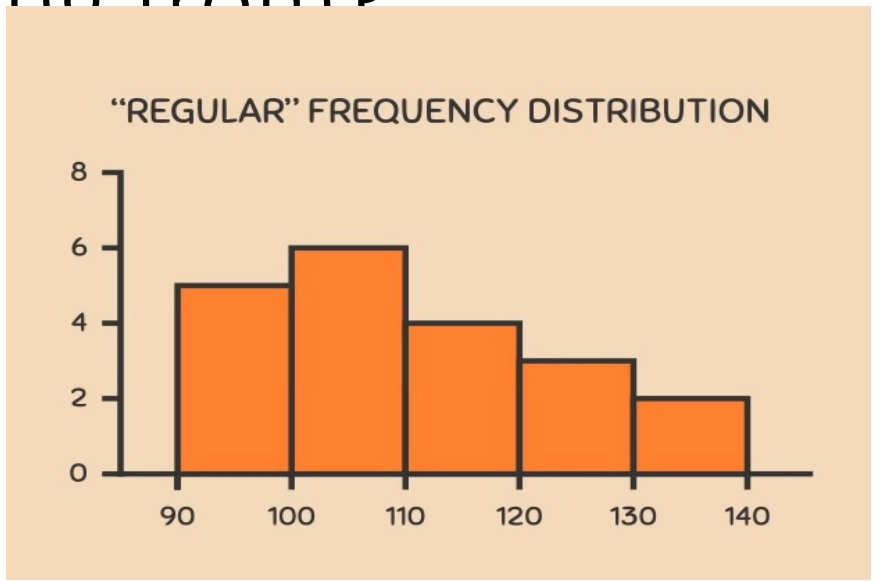
$$P(3; 2) = (0.13534) (8) / 6$$

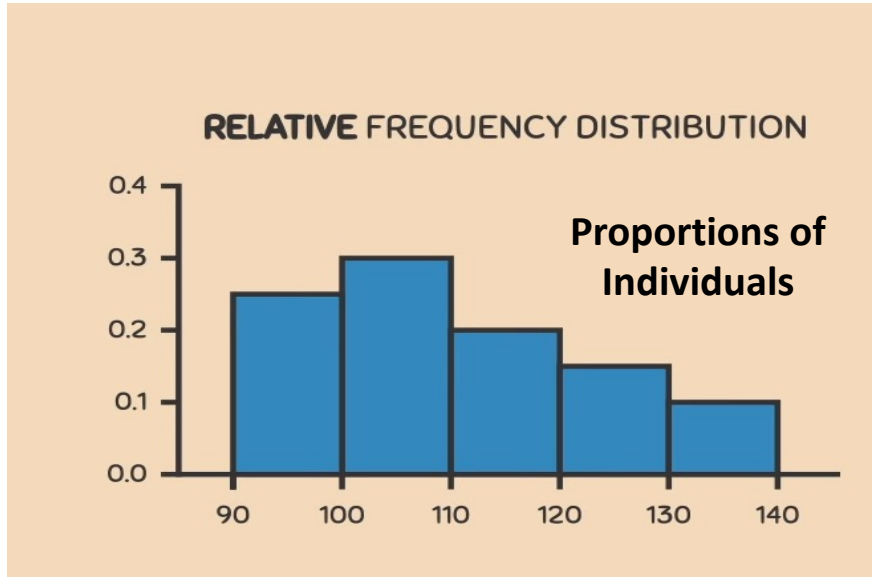
$$P(3; 2) = 0.180$$

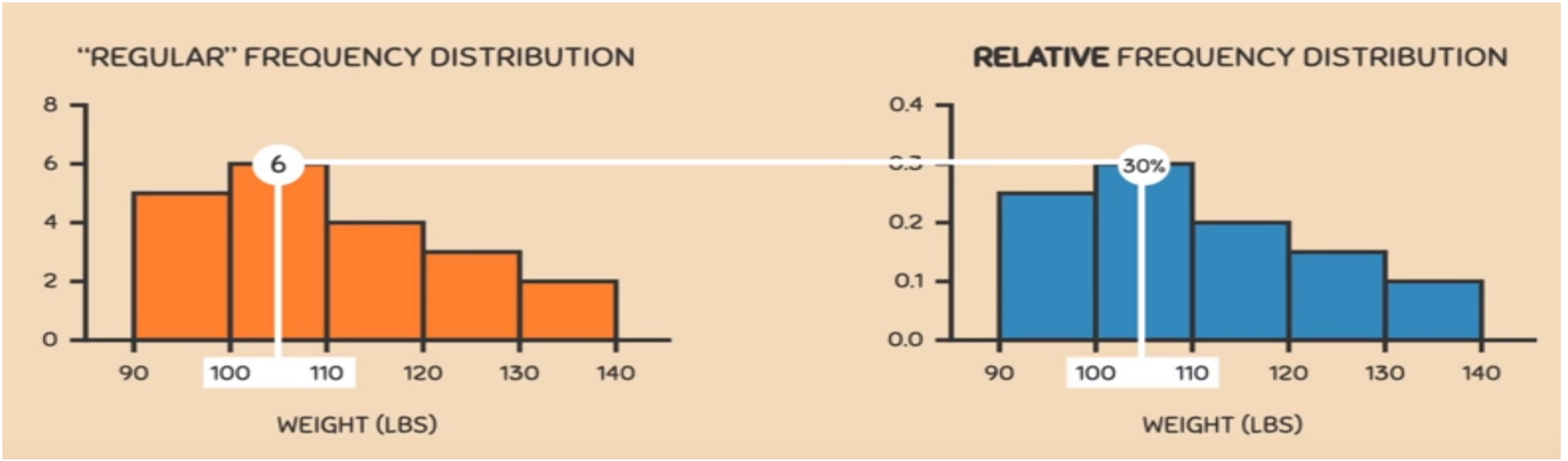
Thus, the probability of selling 3 homes tomorrow is 0.180 .

Density Curves – What are they & Where they are coming from?

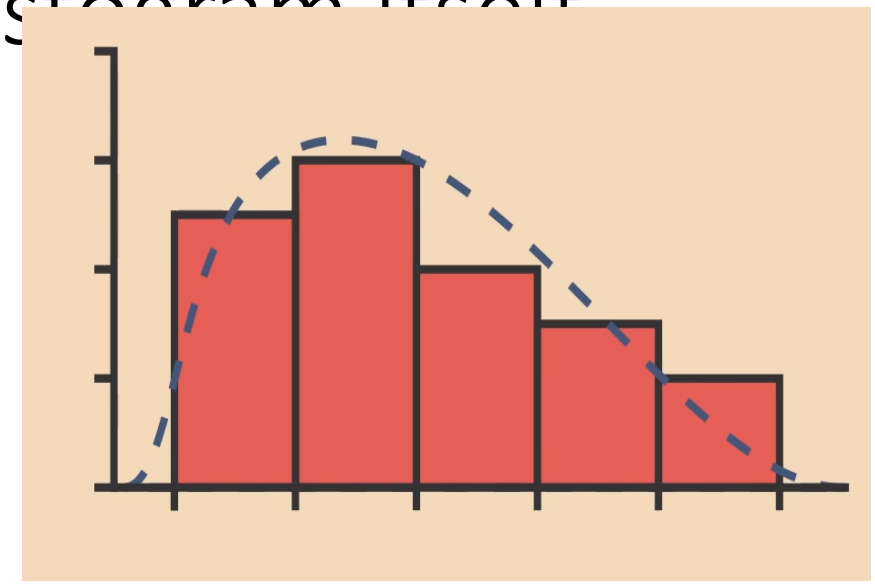
N = 20
Weights in lbs

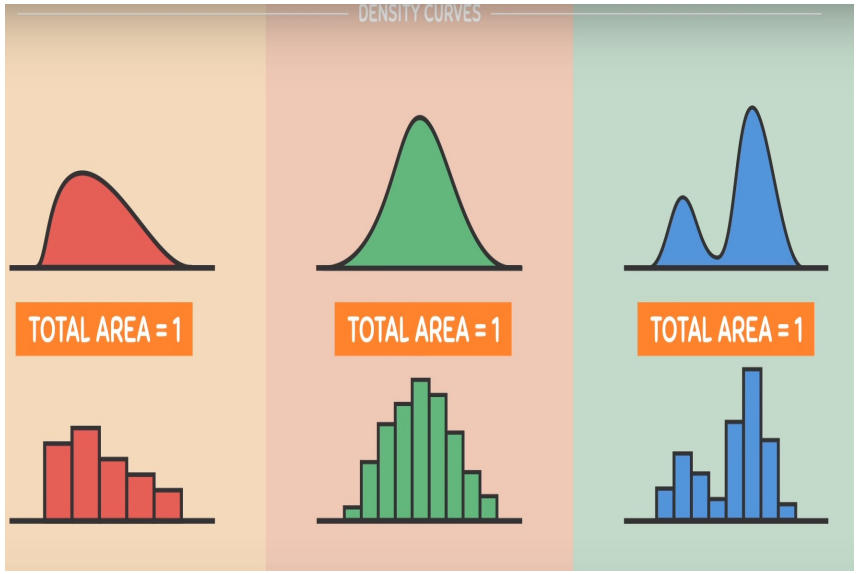


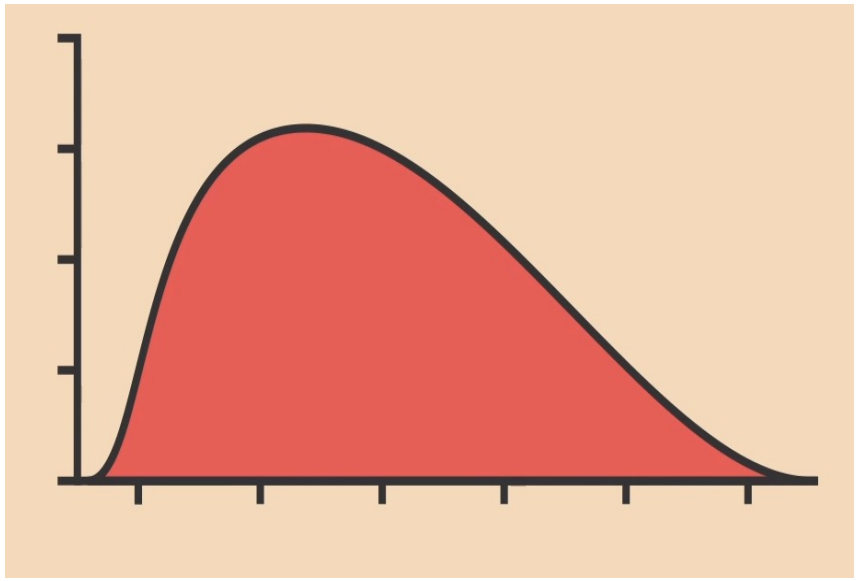




Where it is coming from – A Curve around the histogram itself



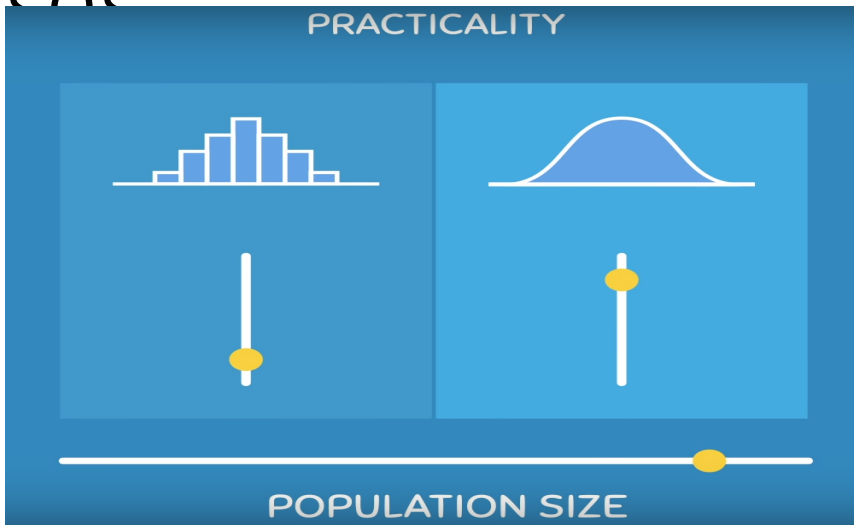


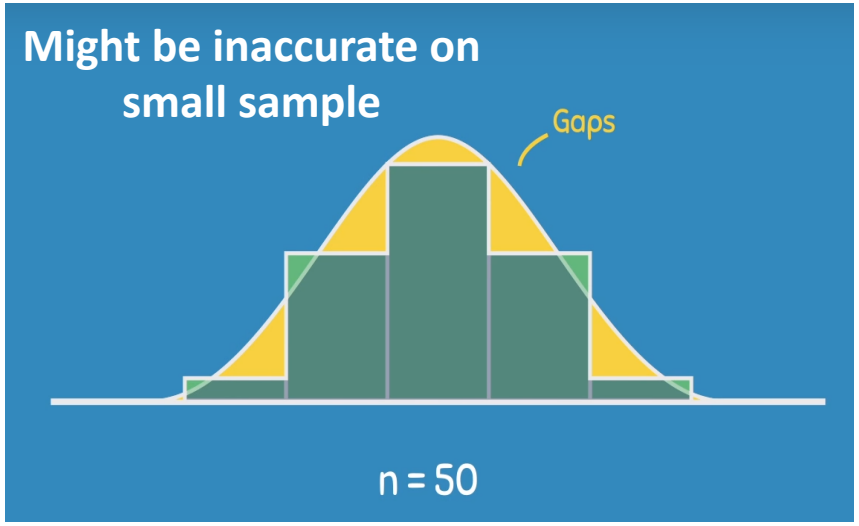


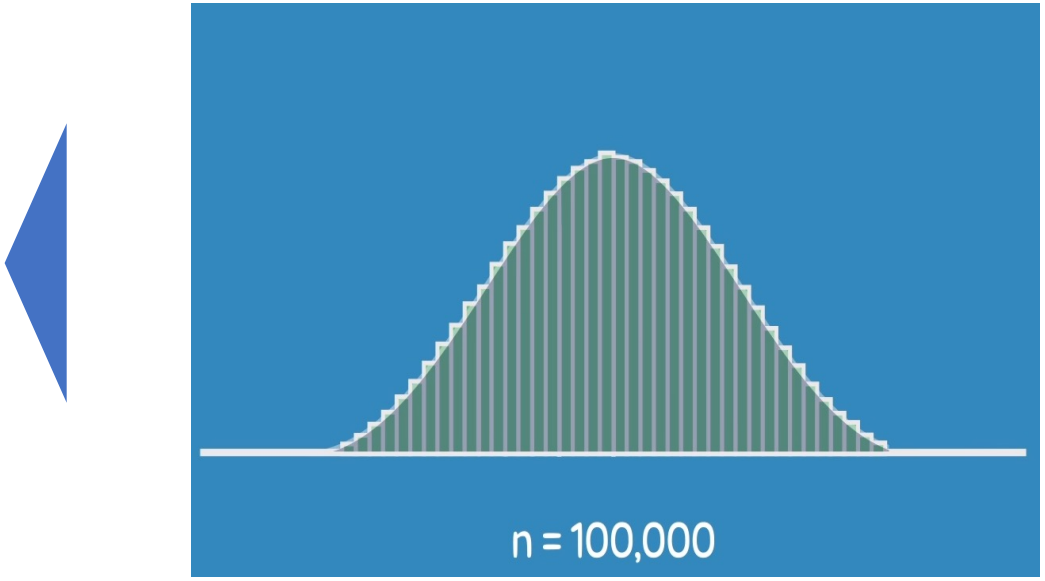
Uses of Histograms

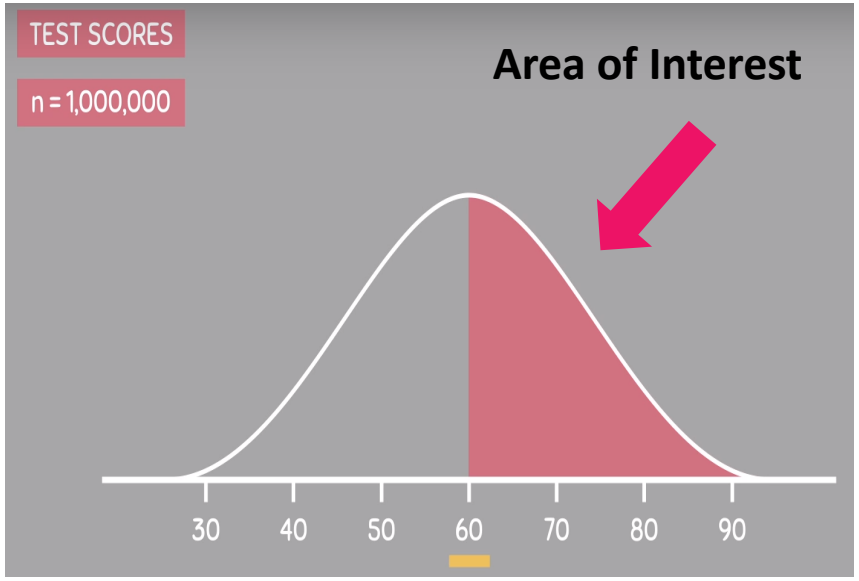
1) **Histograms depends upon how we construct it and depends upon number of intervals we are considering**
2) **Density curves consider infinite amount of intervals – Independent**
3) **Uses of density curves – more popular**

Uses of Density Curves – As Population/Sample size increases

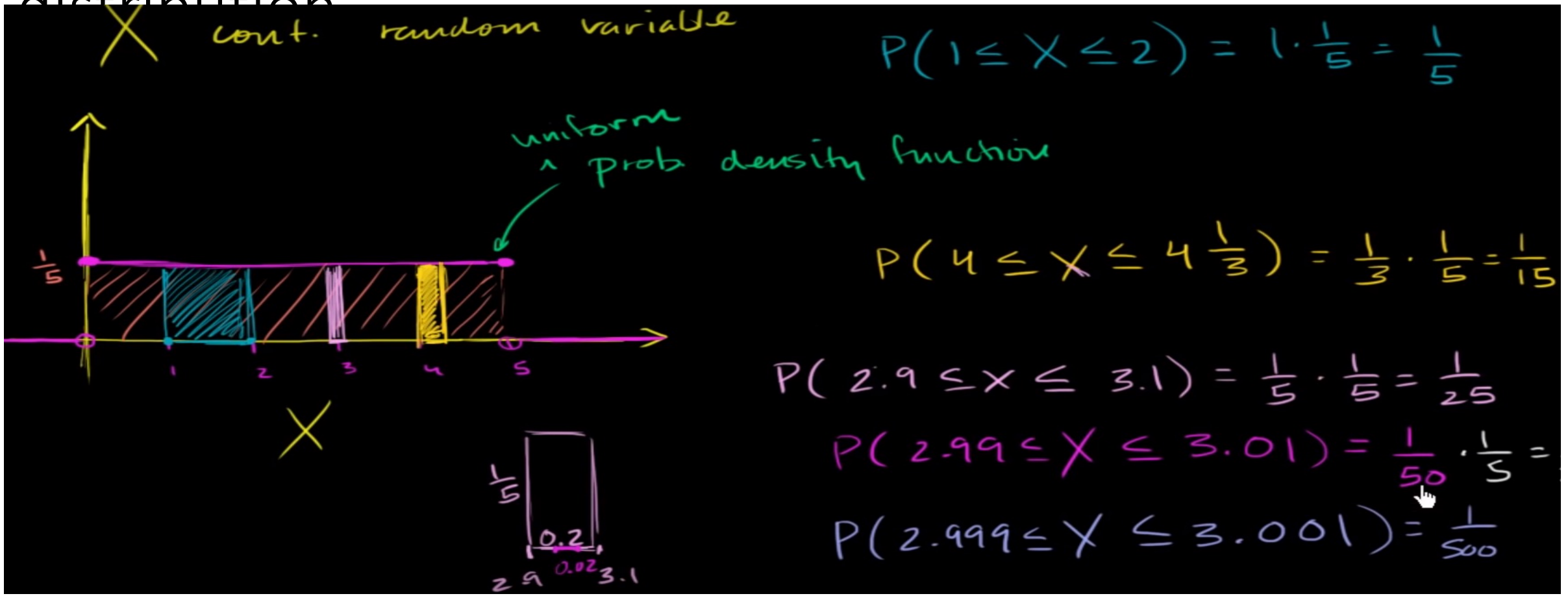


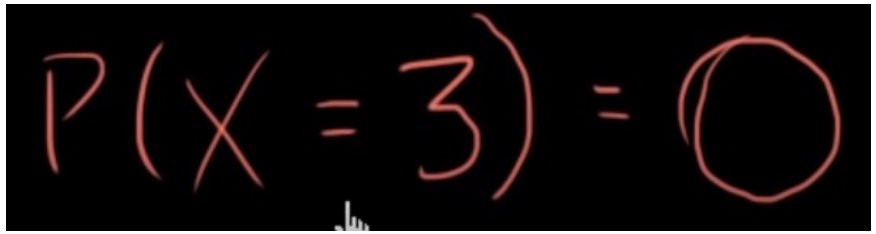






Continuous Probability Distribution Contd. – Area of a distribution



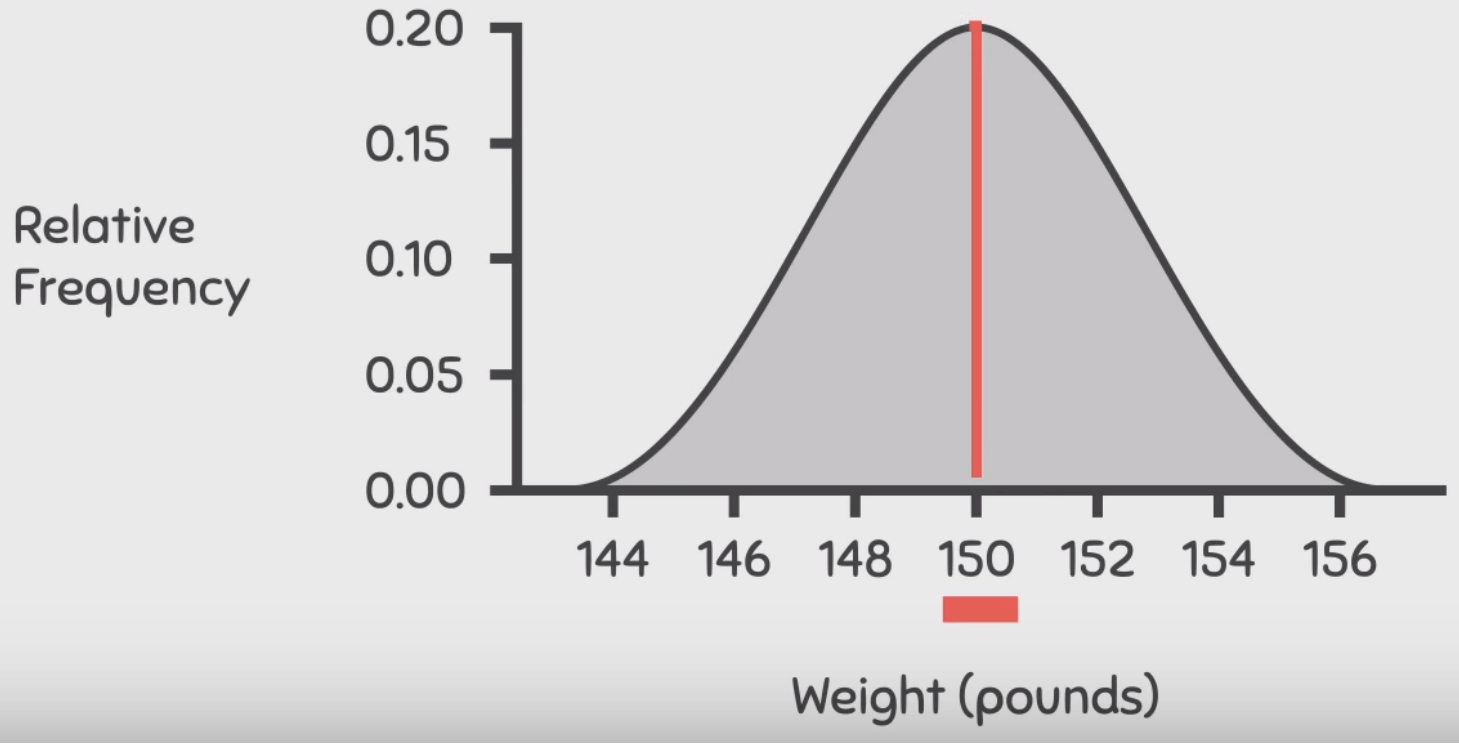


Area = 1 or 100%

Probability of occurrence of any exact value = 0

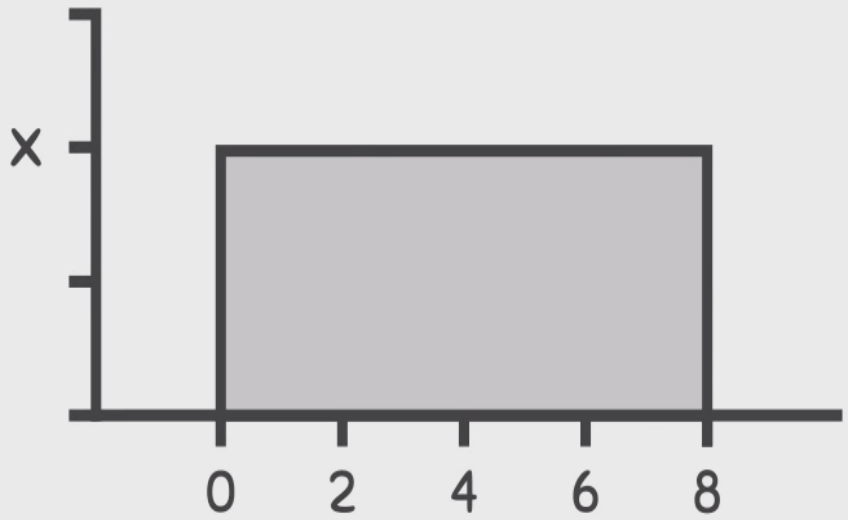
Practice Questions - 1

❶ For the density curve below, approximately what percentage of people weigh exactly 150 pounds?



Practice Questions - 2

❷ For the uniform distribution below, what must be its width in order for it to be a valid density curve?



Area = L x W

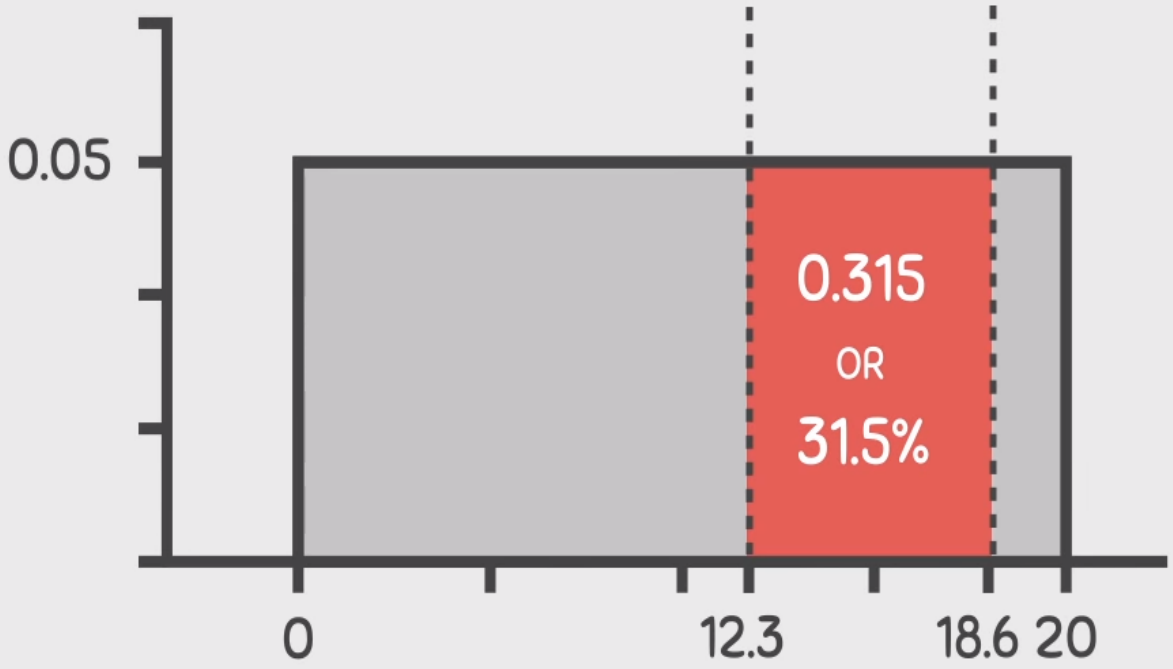
W = Area ÷ L

= 1 ÷ 8

W = 0.125

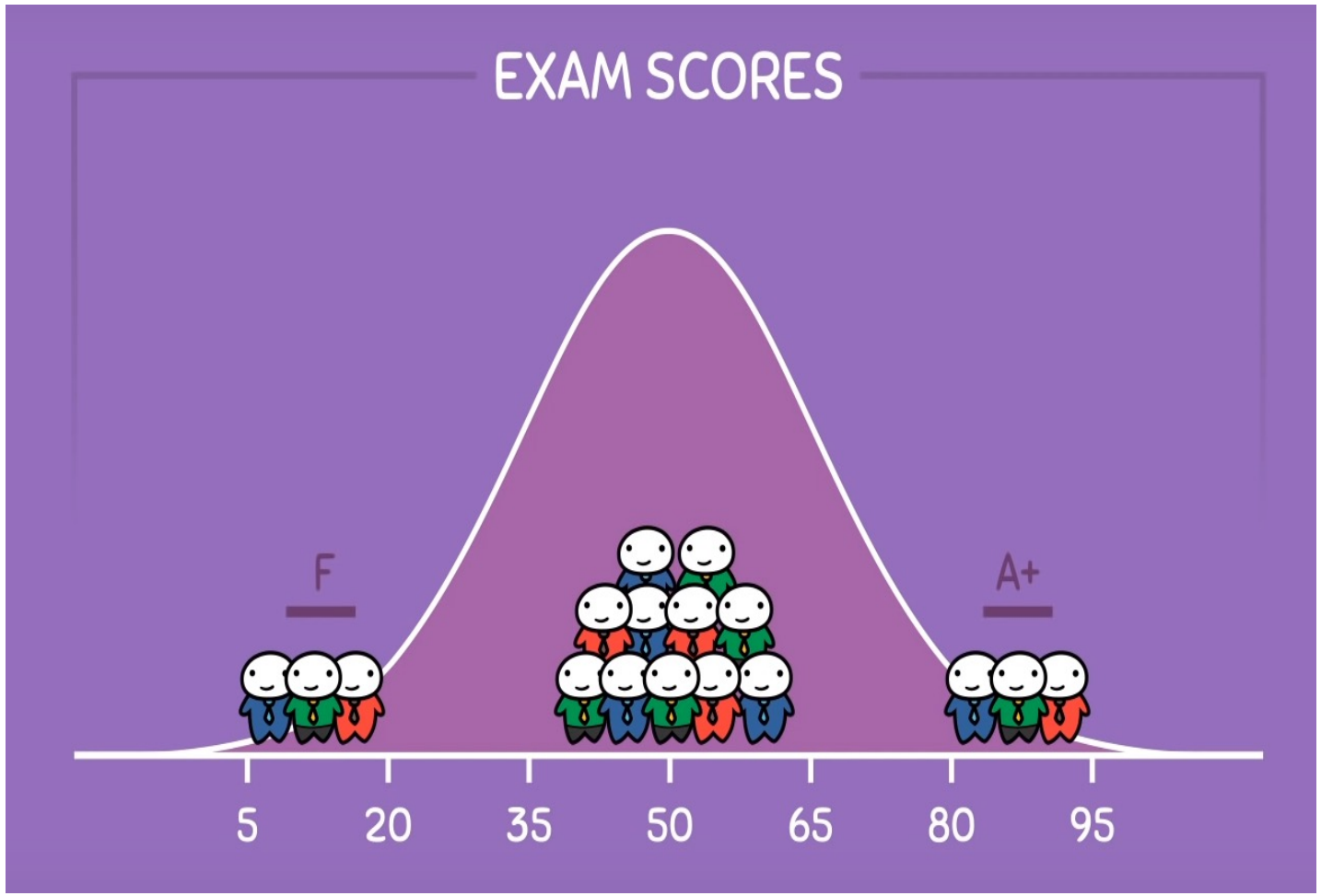
Practice Questions - 3

❸ For the uniform distribution below, what proportion of values are located between 12.3 and 18.6?



Normal Distribution – What it is so practical and common?

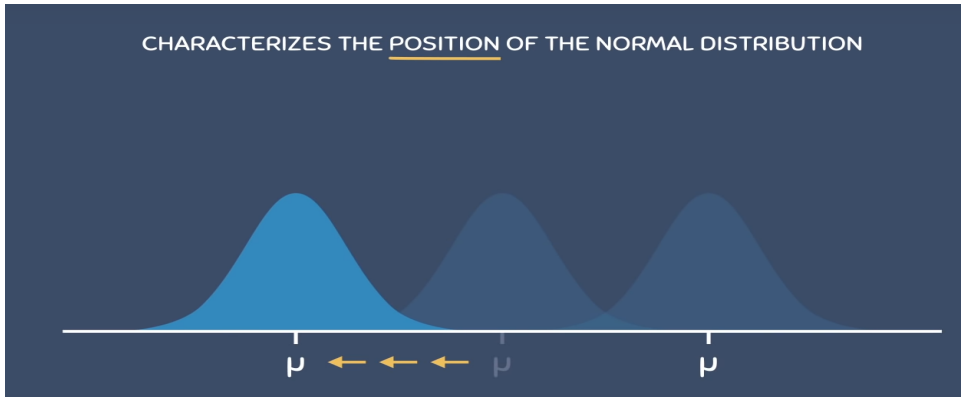
Normal Distribution – A special type of density curve



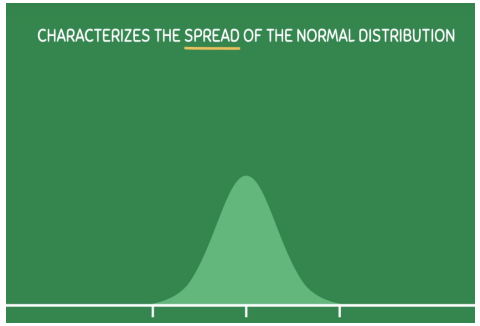
2 parameters μ σ

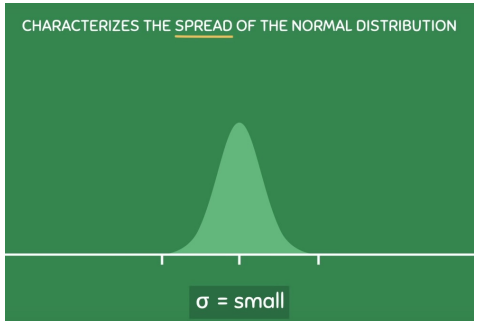
PDF $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

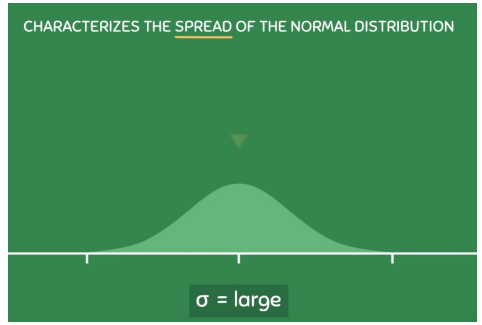
Mean – Position of bell curve



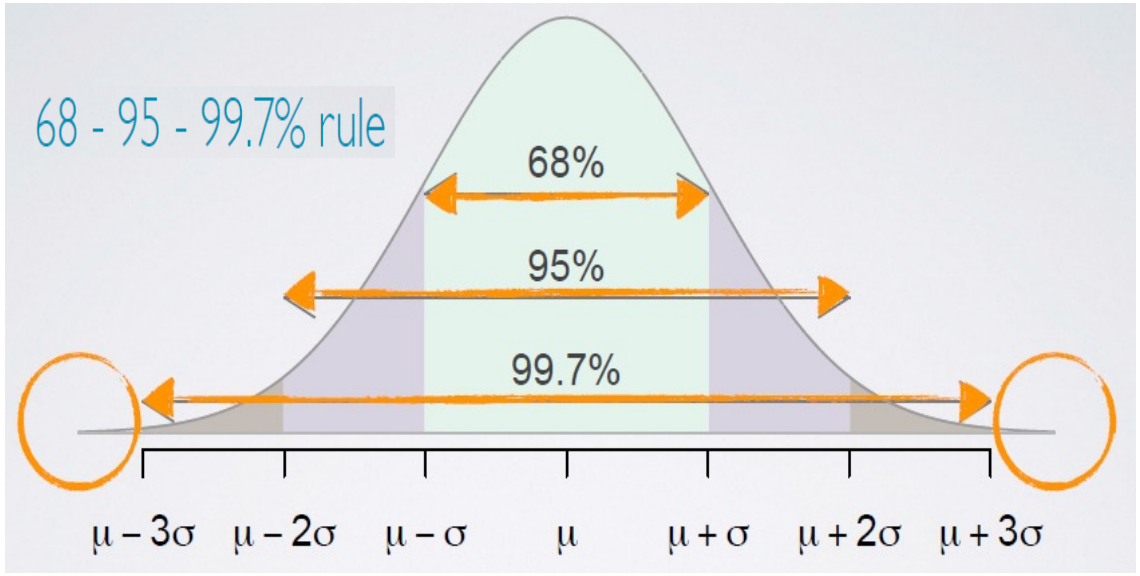
Standard deviation – Shape/Spread of bell curve



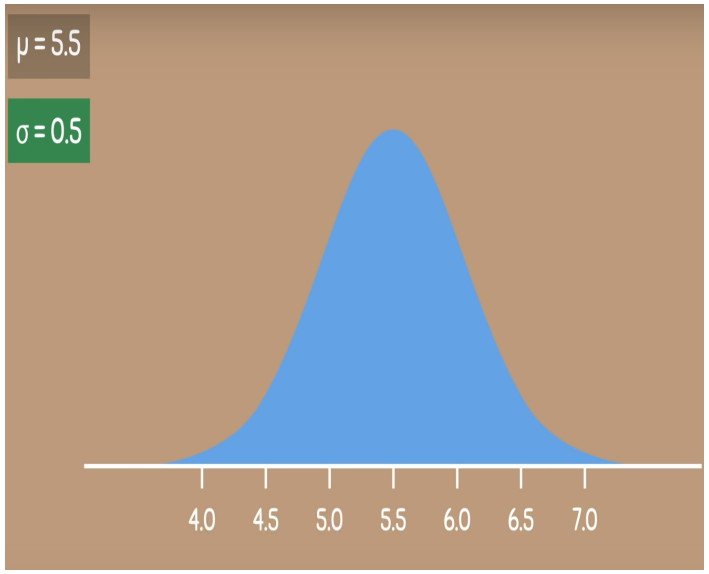


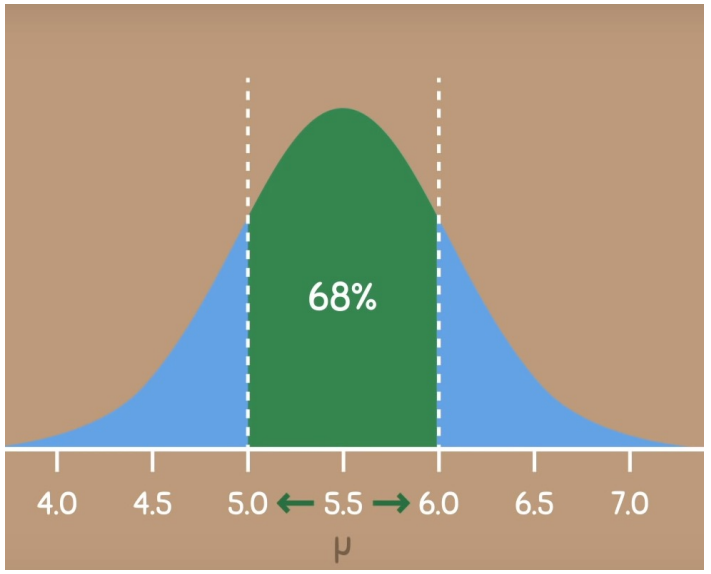


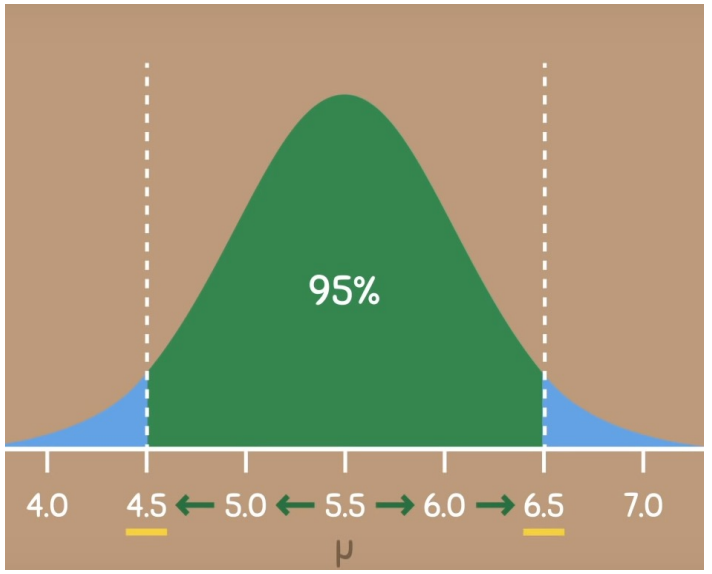
Normal Distribution – the 68-95-99.7 Rule

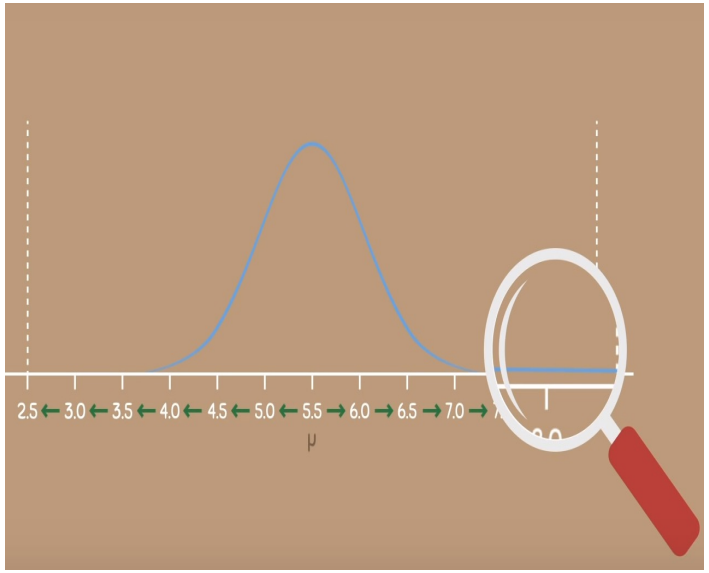


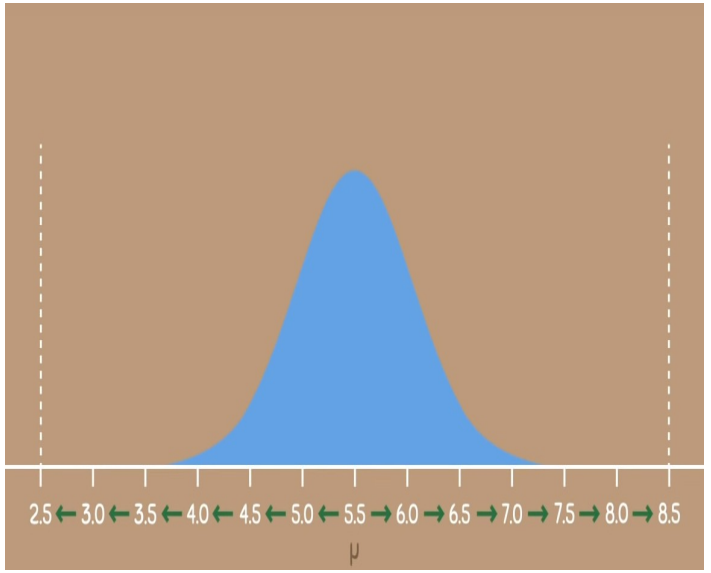
Normal Distribution – the 68-95-99.7 Rule contd..

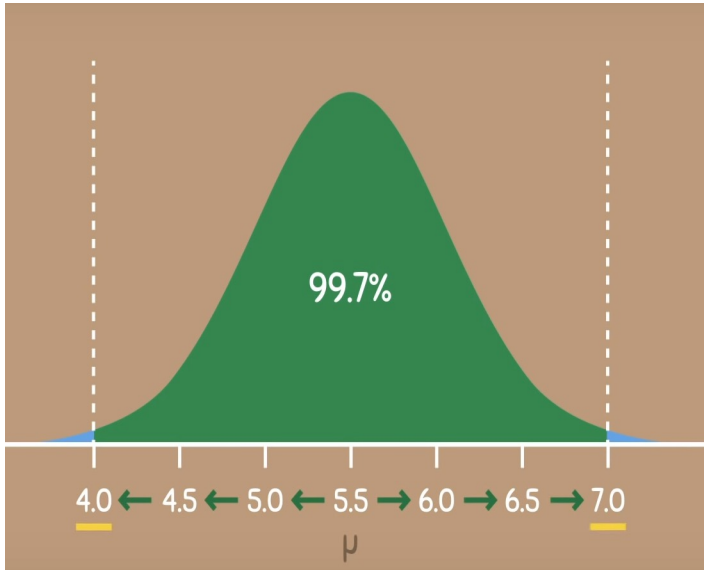










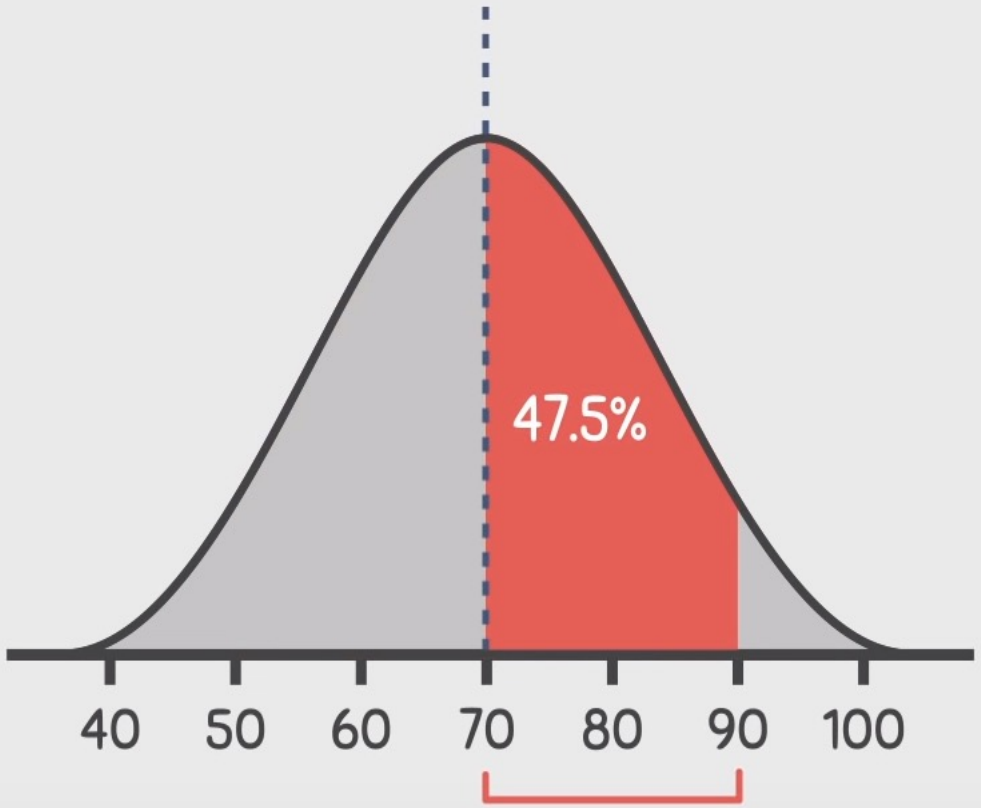


Practice Questions - 1

❶ The normal distribution below has a standard deviation of 10. Approximately what area is contained between 70 and 90?

$\mu = 70$

$\sigma = 10$

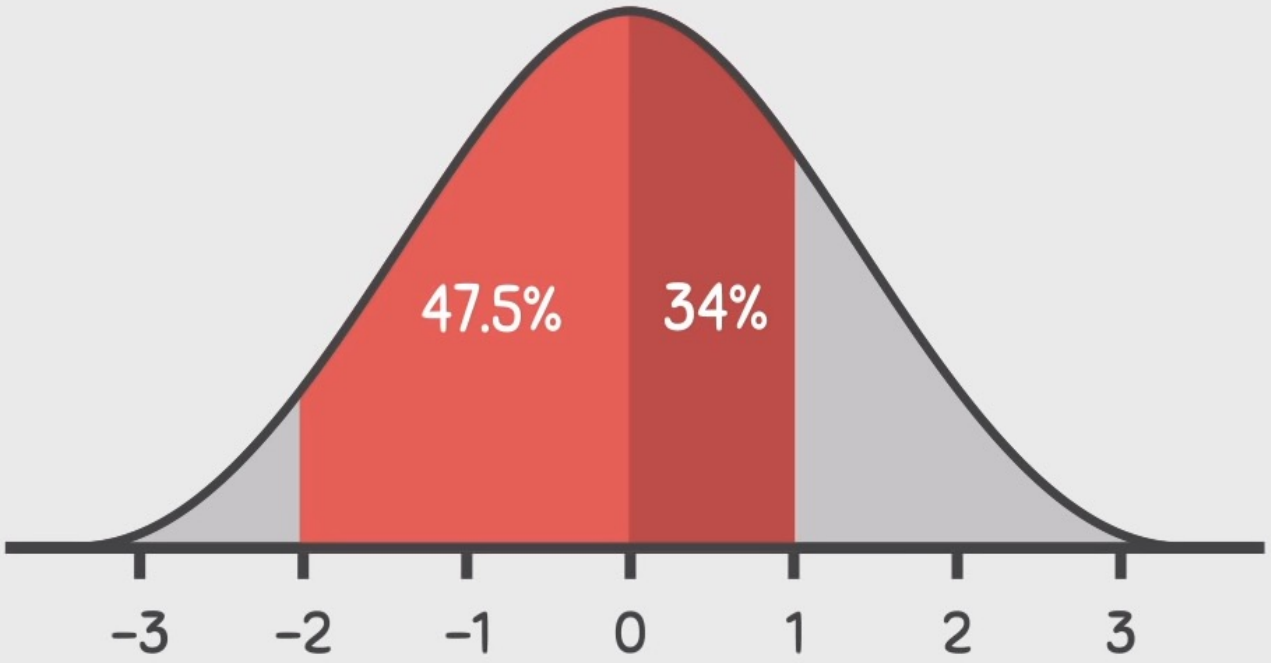


Practice Questions - 1

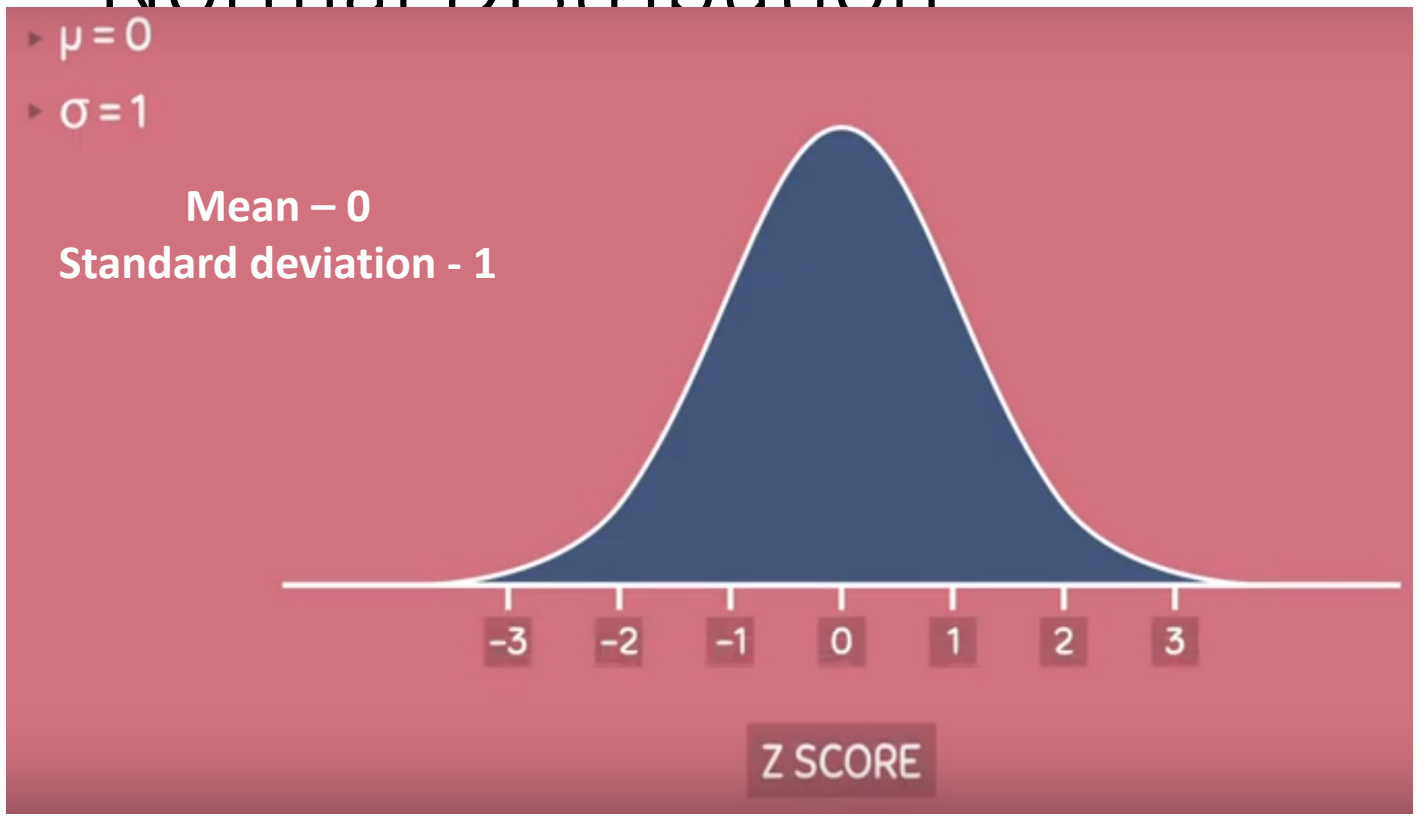
❷ For the normal distribution below, approximately what area is contained between −2 and 1?

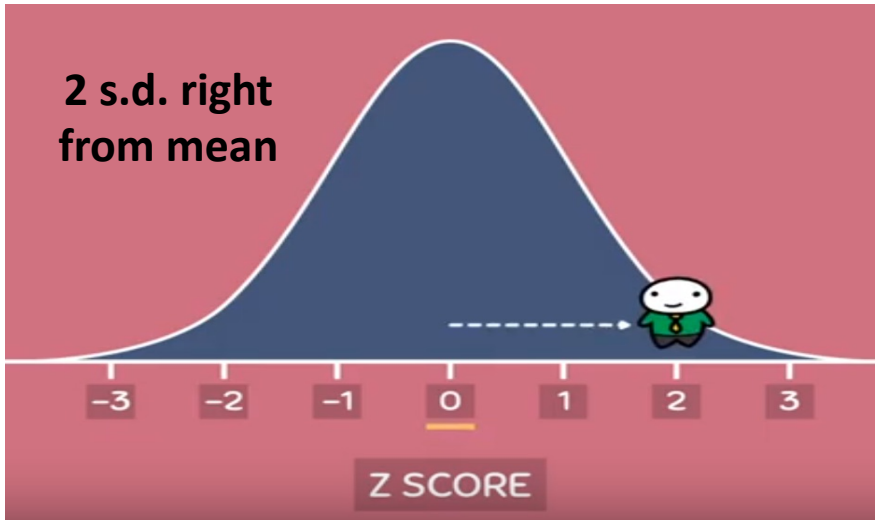
$\mu = 0$

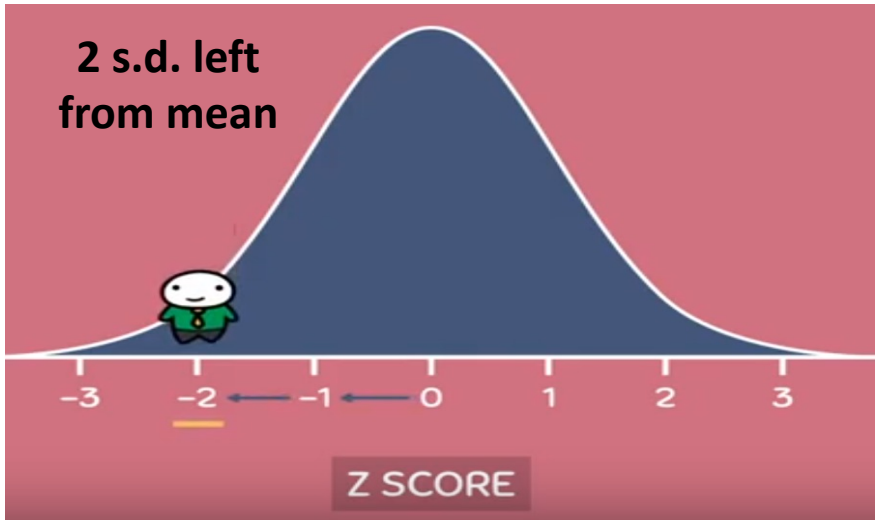
$\sigma = 1$



Standard Normal Distribution – A special type of Normal Distribution





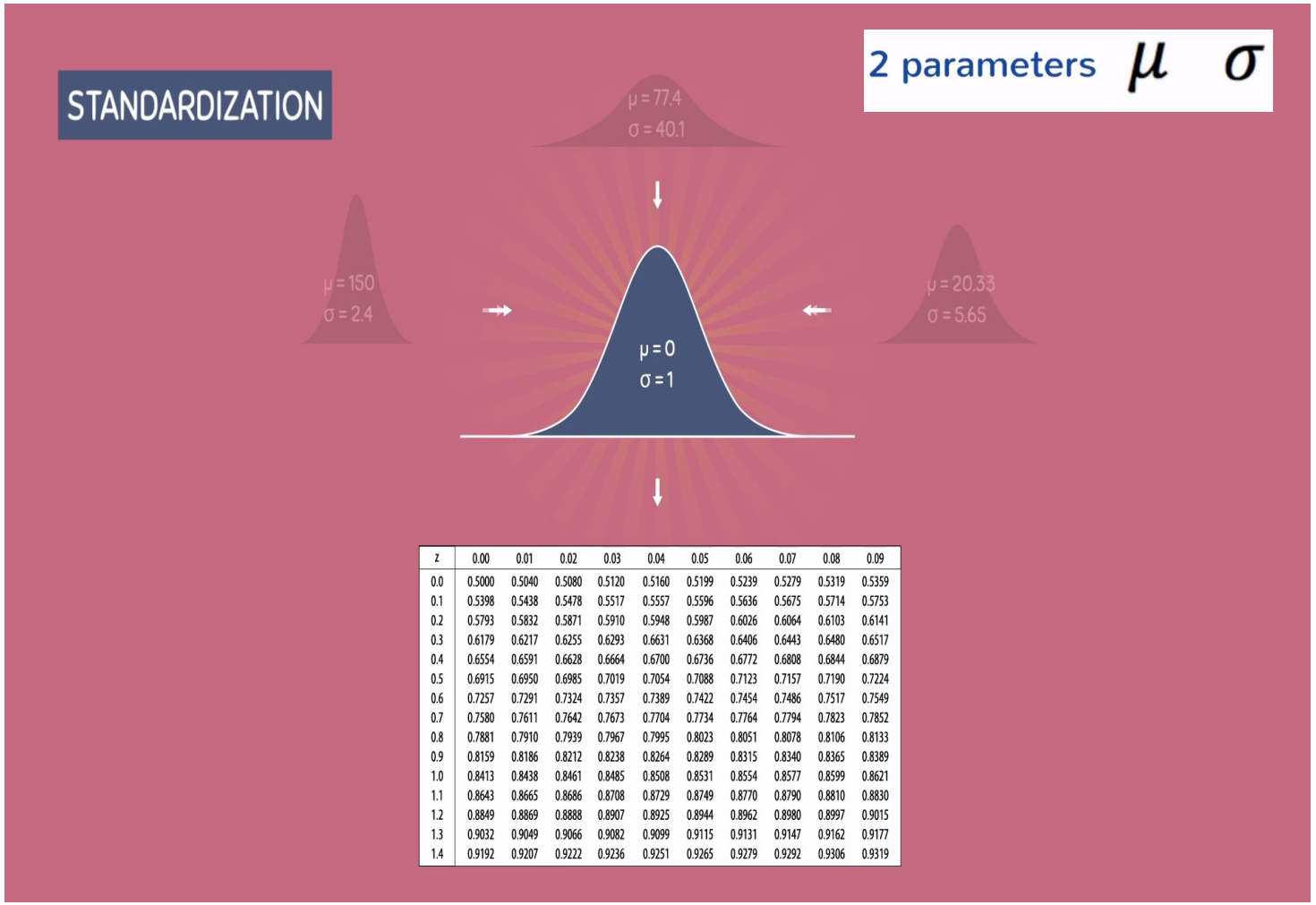


2 parameters μ σ

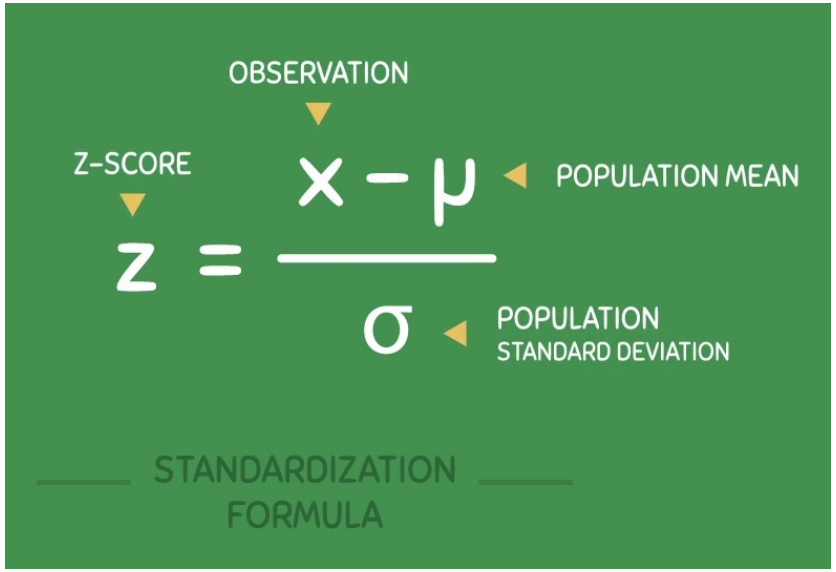
$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Standard Normal Distribution – How & Why do we calculate Z-Scores

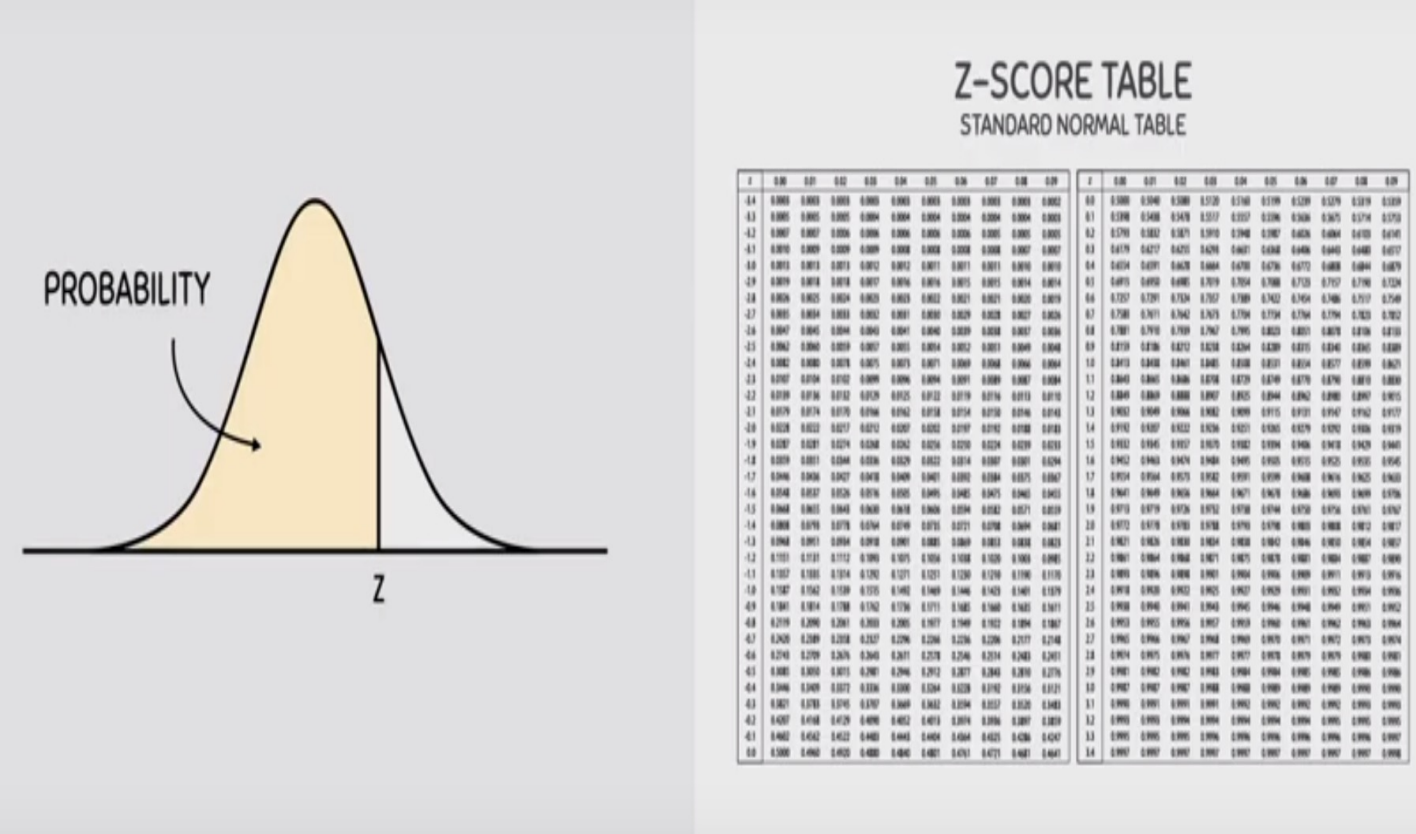
Any Normal distribution can be converted into Standard normal distribution

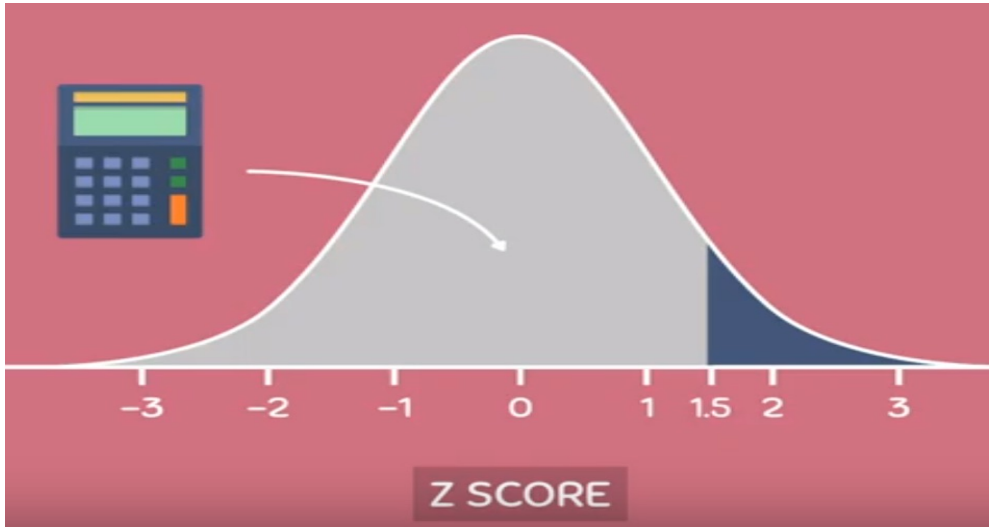


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6631	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319



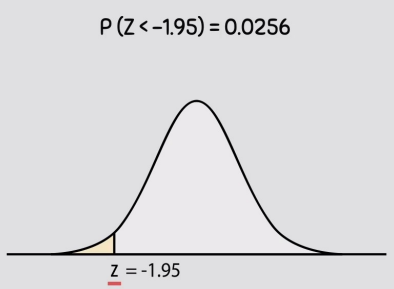
Standard Normal Distribution – Area of Z-scores





2 parameters μ σ

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$



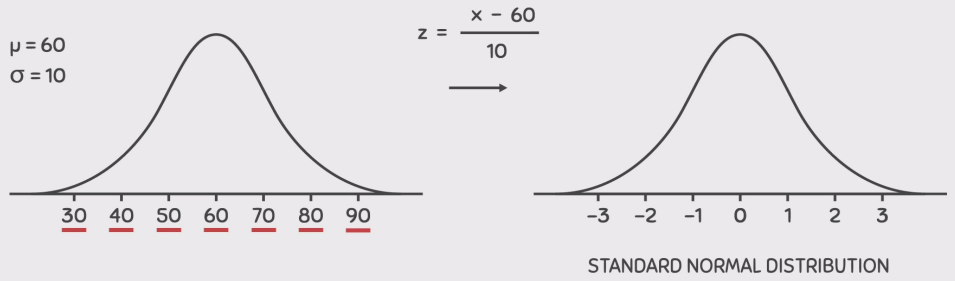
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0224
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660

- **Rows and Columns Z-Scores**
- **Area (percentiles/percentage values inside the tables**

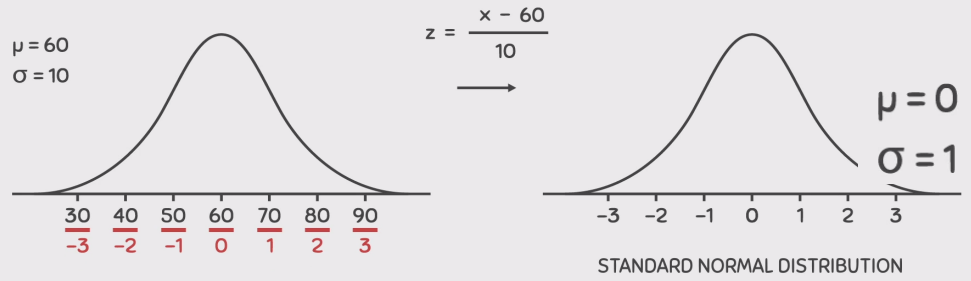
Standard Normal Distribution – How do we calculate Z-Scores contd

Any Normal distribution can be converted into Standard normal distribution

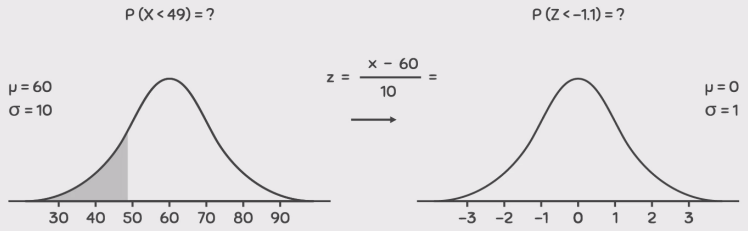
EXAMPLE | Suppose that we gathered data from last year's final chemistry exam and found that it followed a normal distribution with a mean of 60 and a standard deviation of 10.



EXAMPLE | Suppose that we gathered data from last year's final chemistry exam and found that it followed a normal distribution with a mean of 60 and a standard deviation of 10.



EXAMPLE | Suppose that we gathered data from last year's final chemistry exam and found that it followed a normal distribution with a mean of 60 and a standard deviation of 10. What proportion of students scored less than 49 on the exam?



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

Practice Questions - 1

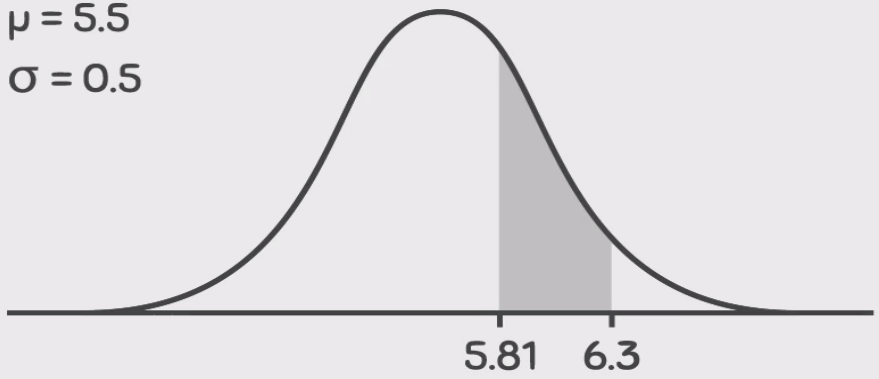
EXAMPLE

When measuring the heights of all students at a local university, it was found that it was normally distributed with a mean height of 5.5 feet, and a standard deviation of 0.5 feet. What proportion of students are between 5.81 feet, and 6.3 feet tall?

$P(5.81 < X < 6.3) = ?$

$\mu = 5.5$

$\sigma = 0.5$



$P(0.62 < Z < 1.6) = 0.2128$

$P(Z < 0.62) = 0.7324$

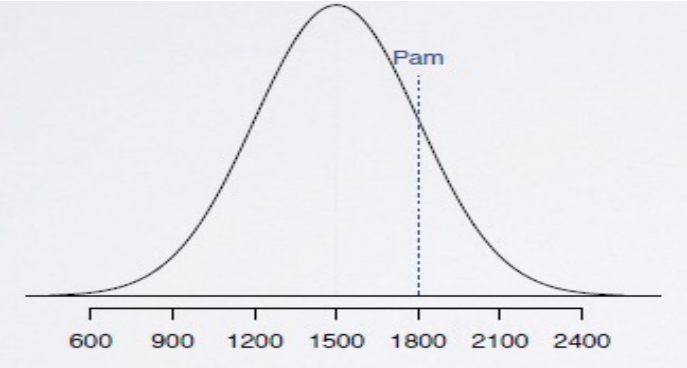
$P(Z < 1.6) = 0.9452$

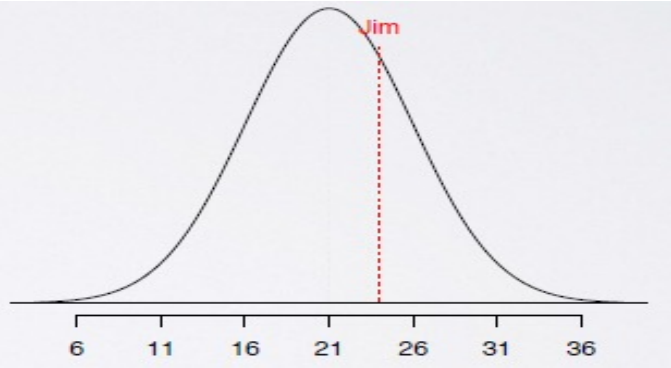
Z-Scores – A more practical use.. To compare two groups

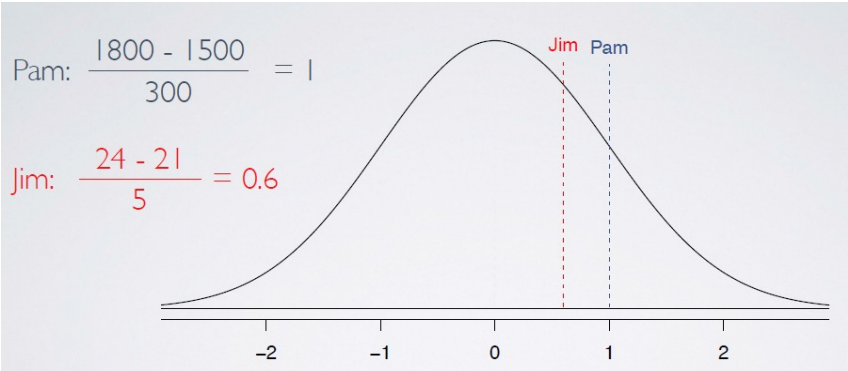
Case:

A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned a 1800 on her SAT, or Jim, who scored a 24 on his ACT?

SAT scores ~ N(mean=1500, SD=300) ; ACT ~ N(mean=21, SD=5)







Key Points :

- Z-score of mean =0
- Unusual observation :|Z| > 2
- Defined for distributions of any shape

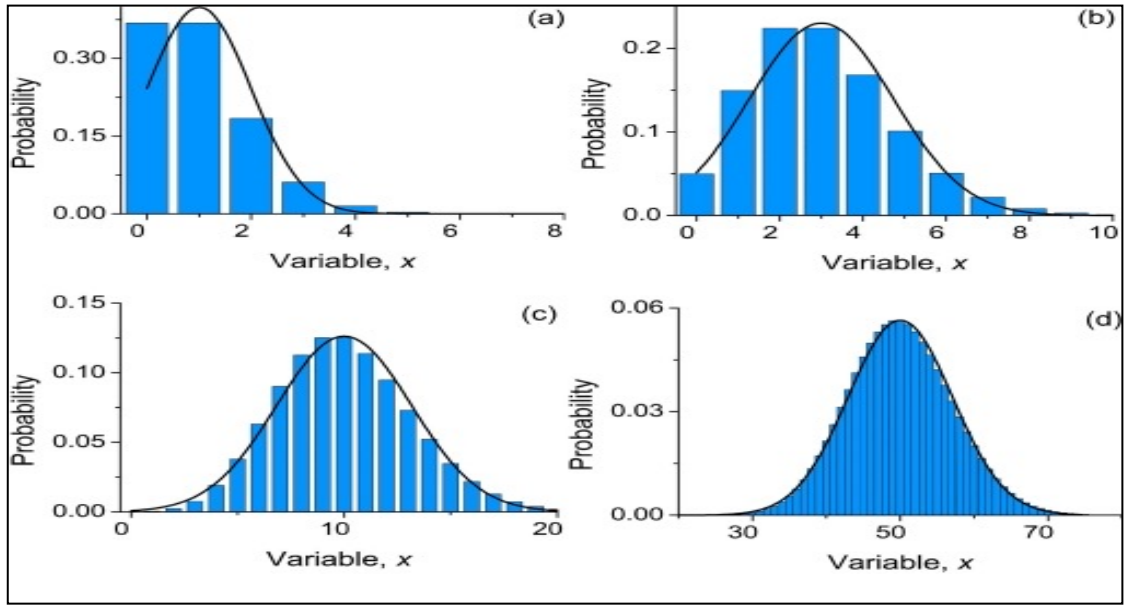
Comparing the SAT and ACT scores using standardized Z-score indicates that indeed Pam has scored better than Jim. This is illustrated by the fact that Pam's scores are further away (higher S.D. of 1) from the mean than Jim's score, thus positioned at a score point that is more than Jim's.

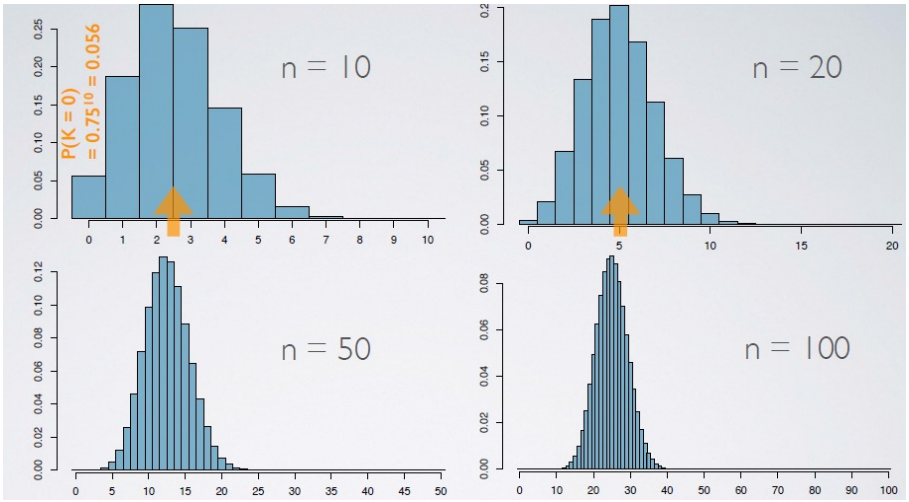
Normal approximation of Binomial and Poisson Distributions

Normal Approximation to Binomial Distribution

With increasing number of trials (‘n’) and probability of success(‘p)’ becoming smaller the skewed Binomial distribution tends to be approximated by a normal distribution. The Success - Failure rule is imperative for the binomial distribution to follow the normal distribution, given as

n> 40 : n*p >=10 and n*P*(1-p) >=10



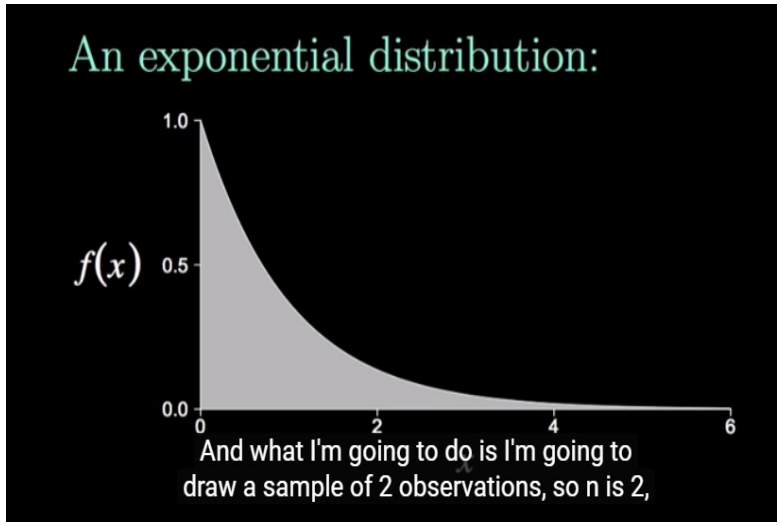


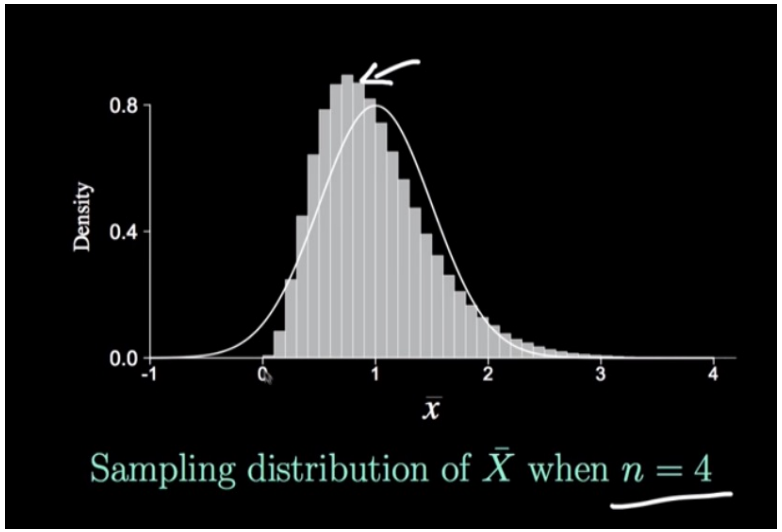
Normal Approximation to Poisson Distribution

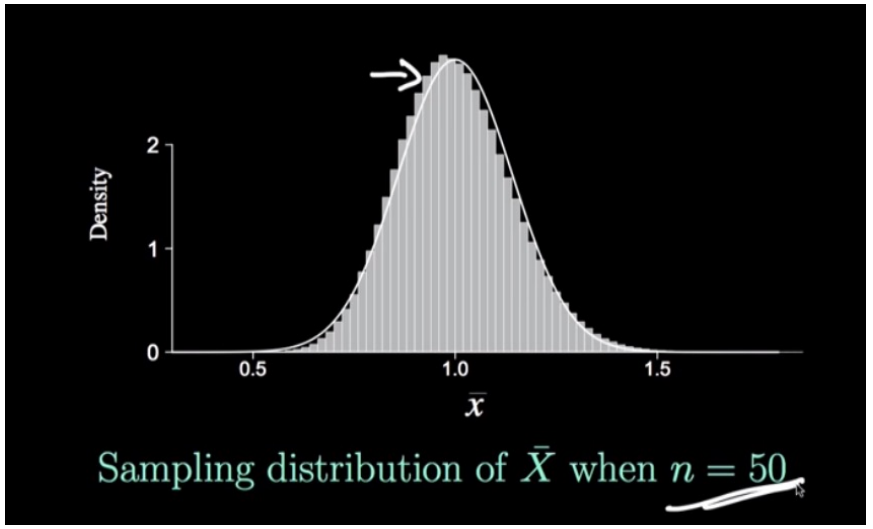
With increasing value of mean(‘m’) the right skewed Poisson distribution tends to be approximated by a normal distribution. The rule for the Poisson to follow a Normal distribution is

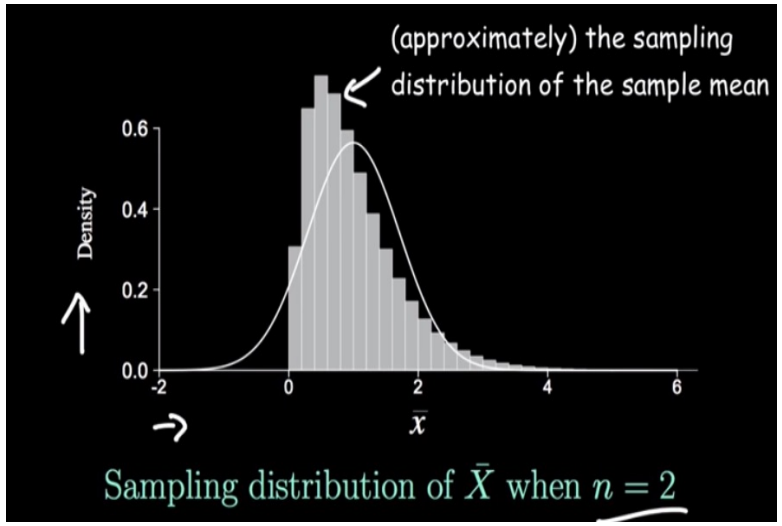
m>10

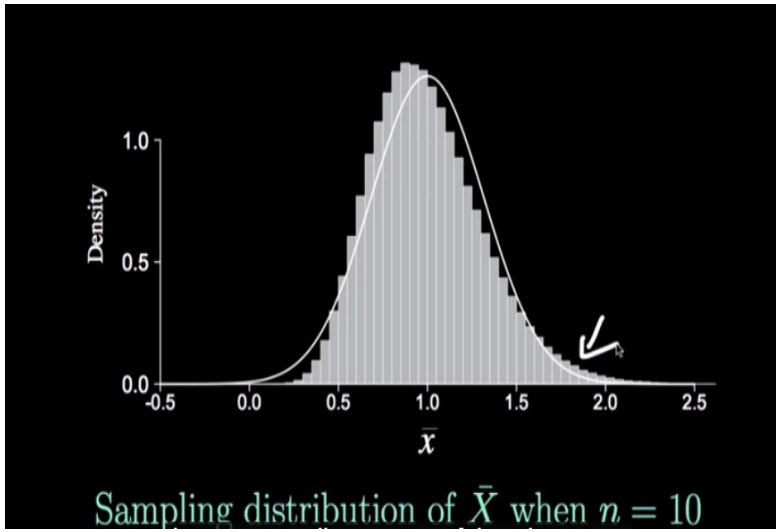
Central Limit Theorem - Visual











As a *very rough guideline*, the sample mean can be considered approximately normally distributed if the sample size is at least 30.

$n \geq 30$

Central Limit Theorem

Simply put, **Central Limit Theorem** states that **the distribution of sample statistics** of a large number of independent, identically distributed variables will be approximately normal, regardless of the underlying distribution

- The mean of the sample means will be the mean of the population
- The variance of the sample means will be the variance of the population divided by the sample size.
- The standard deviation of the sample means (known as the standard error of the mean) will be smaller than the population mean and will be equal to the standard deviation of the population divided by the square root of the sample size.
- If the population has a normal distribution, then the sample means will have a normal distribution.
- If the population is not normally distributed, but the sample size is sufficiently large, then the sample means will have an approximately normal distribution

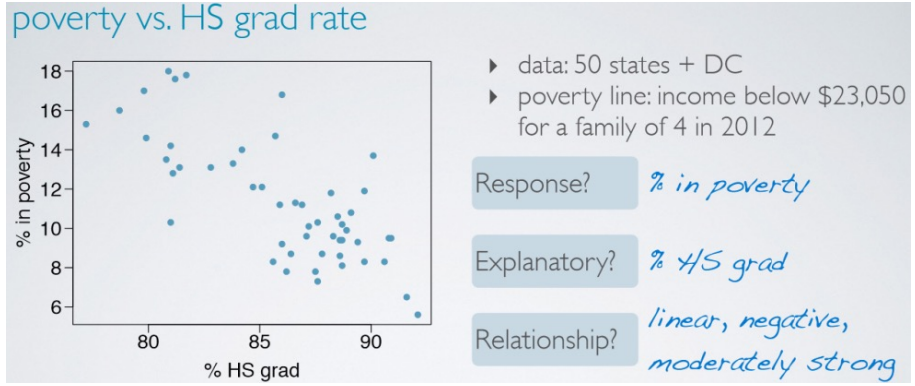
Assumptions:

- **Independence:** Sampled observations must be independent.
- **Large sample size in case of non normal distribution**

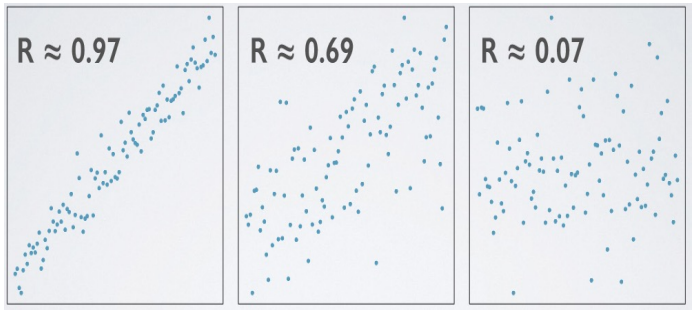
Why is the Central Limit Theorem Important?

If we know the population mean and standard deviation, we know the following will be true:
The distribution of means across repeated samples will be normal with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of n. Since we know exactly what the distribution of means will look like for a given population, we can take the mean from a single sample and compare it to the sampling distribution to assess the likelihood that our sample comes from the same population.

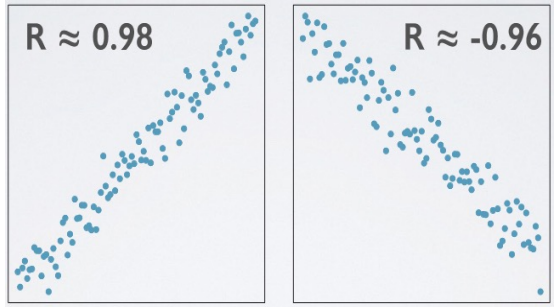
Correlation Concept



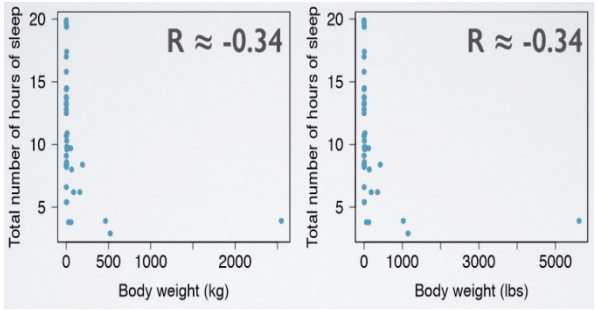
Correlation describes the strength of the *Linear Association* between two Variables. The Correlation Co-efficient is denoted by r which is always in between -1 (perfect linear negative relationship) and 1 (perfect linear positive relationship)



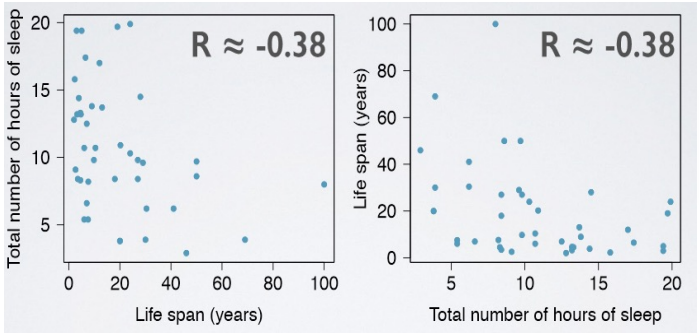
Absolute value measures strength



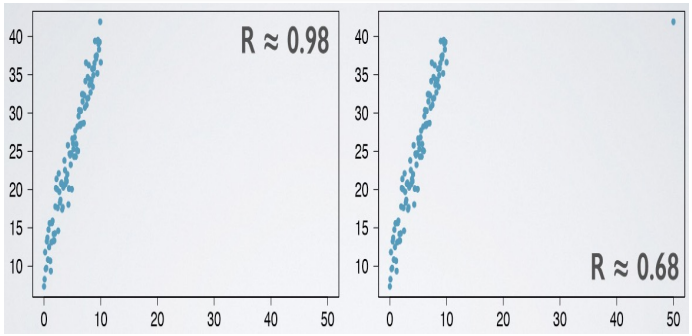
Sign indicates Direction



Not Affected by Unit Change

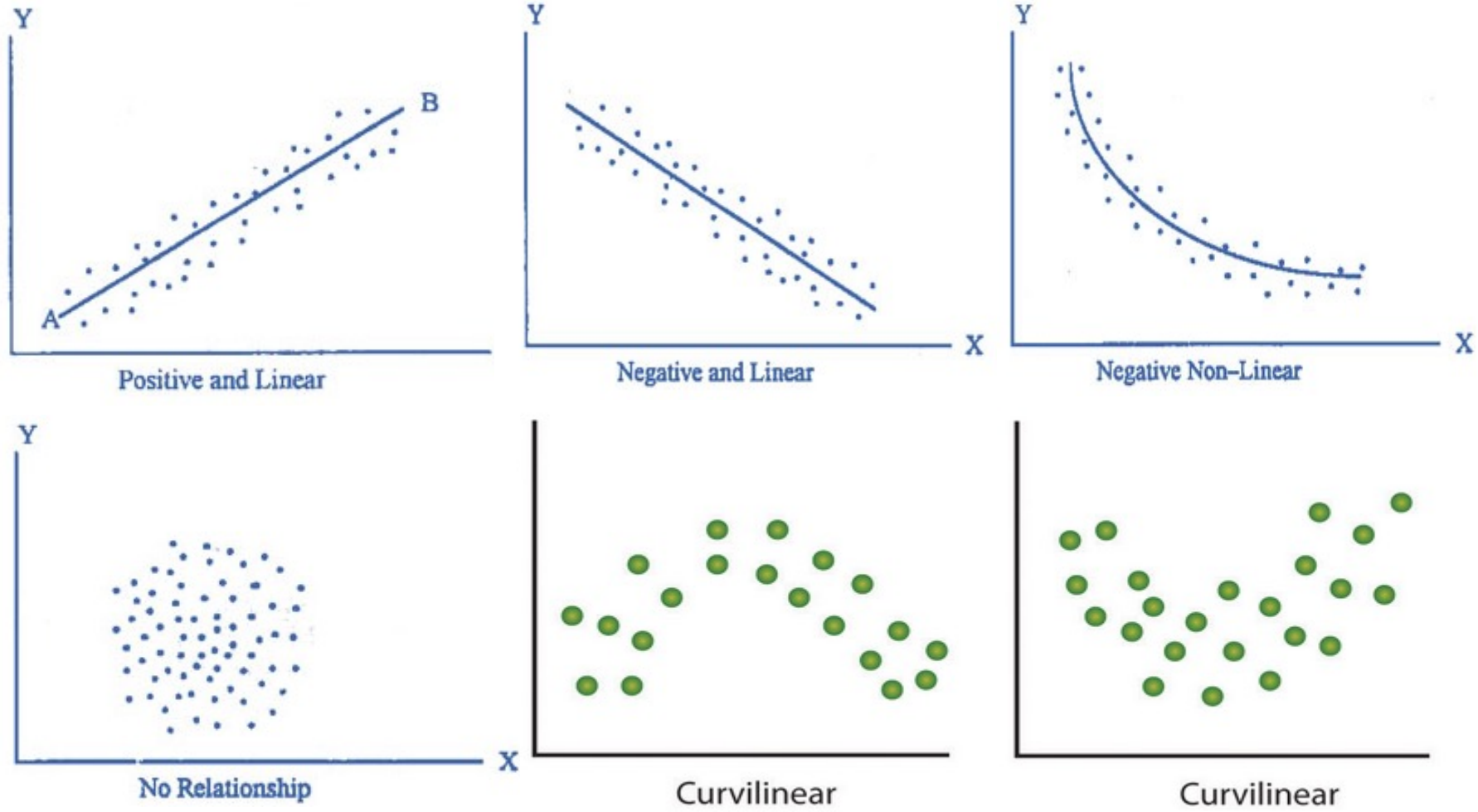


Corr.(Y and X) = Corr.(X and Y)



Outlier Sensitive

Scatter Plots…Guess the Correlation Coefficient…!!!



Pearson Correlation…

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship

- *Limit: Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists..*
- *Pure number: It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.*
- *Symmetric: Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.*

Degree of correlation:

- *Perfect: If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).*
- *High degree: If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.*
- *Moderate degree: If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.*
- *Low degree: When the value lies below + .29, then it is said to be a small correlation.*
- *No correlation: When the value is zero.*

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since

$$\operatorname{cov}(X,Y) = \operatorname{E}[(X-\mu_X)(Y-\mu_Y)],$$[7]

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\operatorname{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \quad \text{(Eq.2)}$$

where:

- σ_Y and σ_X are defined as above
- μ_X is the mean of X
- μ_Y is the mean of Y
- E is the expectation.

Rank Correlation…

	Marks									
English	56	75	45	71	61	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i = difference in paired ranks and n = number of cases.

Rank Correlation: Quick Maths…

Example 9 : 39 In a contest, two judges ranked seven candidates in order of their preference as in the following table :

Candidates		A	B	C	D	E	F	G
Ranks by Judge	I	2	1	4	5	3	7	6
,, ,, ,,	II	3	4	2	5	1	6	7

Calculate the rank correlation coefficient.

Solution :

TABLE 9.17—**Calculations for Rank Correlation Coefficient**

Candidates	*Ranks by* Judge I x	Judge II y	$d=x-y$	d^2
A	2	3	—1	1
B	1	4	—3	9
C	4	2	2	4
D	5	5	0	0
E	3	1	2	4
F	7	6	1	1
G	6	7	—1	1
Total	—	—	0	20

Here, $n = 7$, $\Sigma d^2 = 20$,

$$\therefore \quad R = 1 - \frac{6\,\Sigma d^2}{(n^3 - n)} = 1 - \frac{6 \times 20}{(7^3 - 7)} = 0.64.$$

Ans. $R = 0.64$

Regression Analysis…

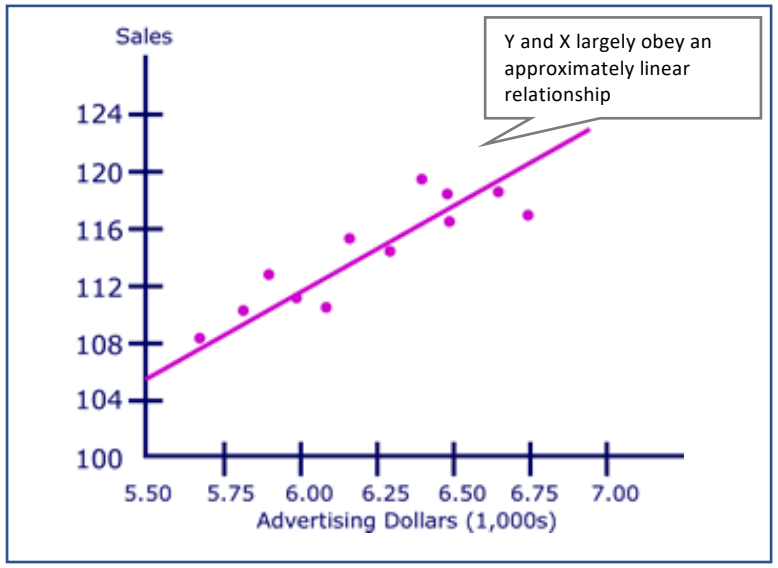
- ***Regression analysis*** is a statistical technique used to describe relationships among variables.
- The simplest case to examine is one in which a variable **Y** , referred to as the ***dependent*** or ***target*** variable, may be related to one variable **X**, called an ***independent*** and ***explanatory*** variable , or simply a ***regressor***.
- If the relationship between **Y** and **X** is believed to be linear, then the equation for a line may be appropriate:

  $\mathbf{Y} = = \beta 1 + \beta 2\ \mathbf{X}$,

  Where $\beta 1$ is an intercept term and $\beta 2$ is a slope coefficient
- This is an exact or ***deterministic*** relationship
- Deterministic relationships are very rarely encountered in business environments

  Assets = Liabilities + Owner equity total costs

  = Fixed cost + variable costs



How to fit a line to describe the "broadly linear" relationship between Y and X when the (x, y) pairs do not all lie on a straight line?

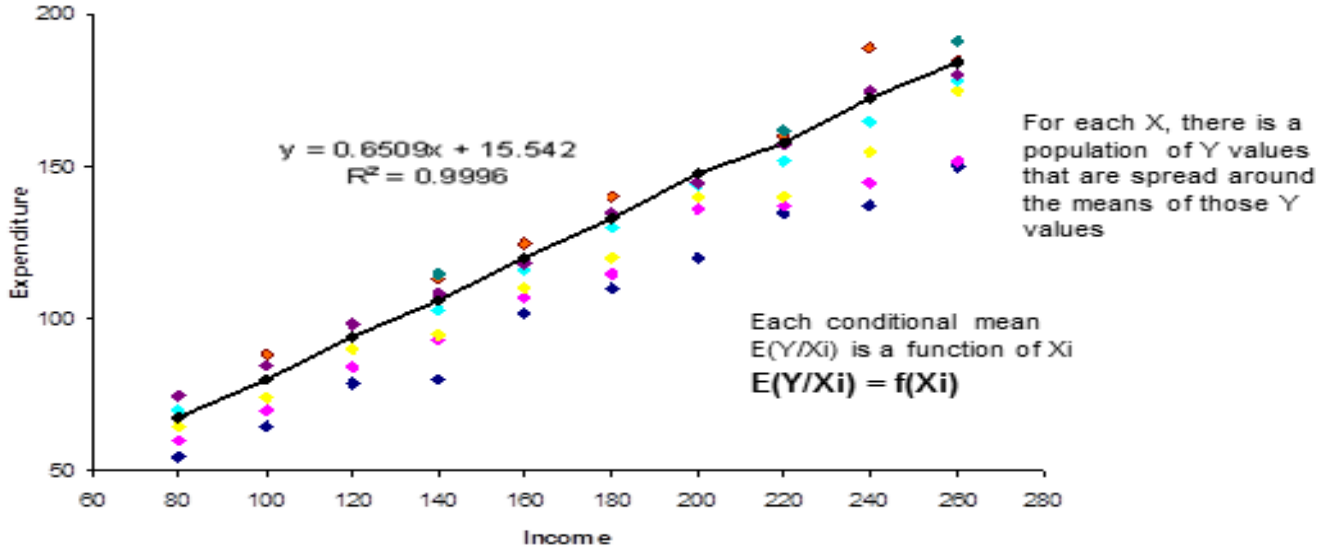
- Consider the pairs (xi,yi)
- Let ˆyi be the **predicted** value of yi
- ei = yi- ˆyi as the **residual** representing the **error** involved
- Several approaches are there to minimize the error :
  1. Minimize the sum of the errors
  2. Minimize the sum of the absolute error
  3. ***Least square method***

Regression Analysis…

Lets take a data of 60 families with their weekly income & weekly consumption expenditure.

The 60 families are divided into 10 income groups and the weekly consumption expenditure of each family in each income group is shown below:

Data	Weekly Income									
Y(Exp) / X(Income)	80	100	120	140	160	180	200	220	240	260
Weekly Expenditure	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	-	88	-	113	125	140	-	160	189	185
	-	-	-	115	-	-	-	162	-	191
Conditional Mean of Y, E(Y/X)	67.5	80.29	94.17	105.9	119.7	132.9	147.5	157.9	172.3	183.9

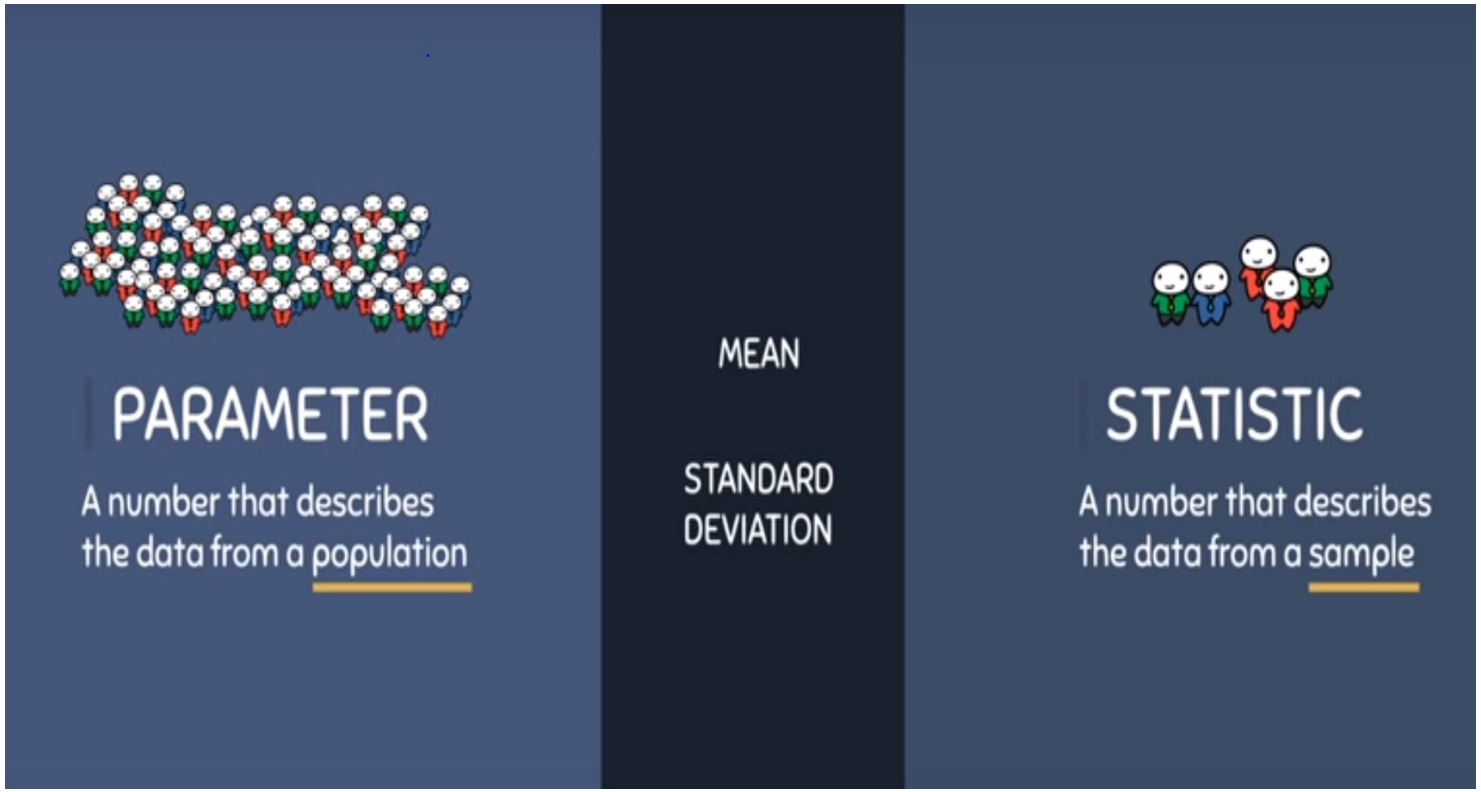


There is considerable variation in the weekly consumption expenditure in each income group. The general picture is that, despite the variability, on an average the weekly consumption expenditure increases as income increases.

In the above case, we see that E(Y/Xi) is a linear function of Xi (i.e. expenditure is linearly related to income)

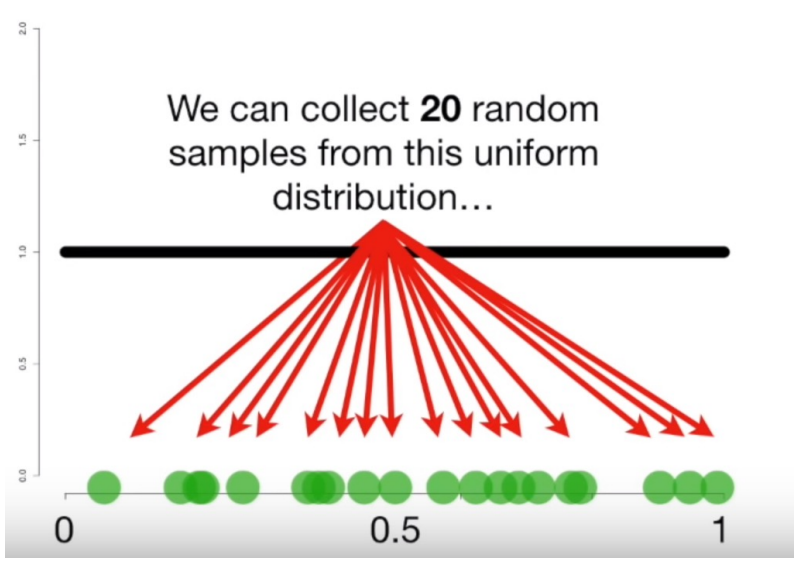
E(Y/Xi) = β1 + β2Xi (where β1 & β2 are also known as intercept and the slope coefficients respectively)

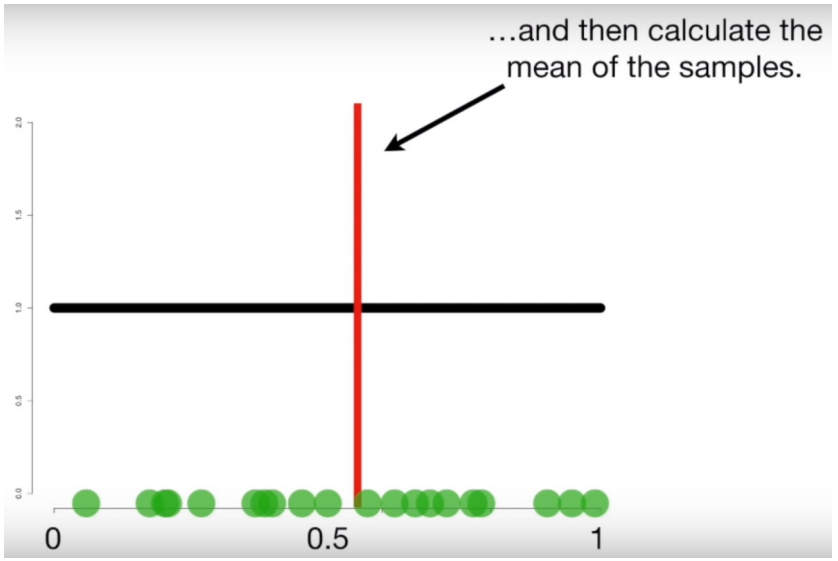
Parameter v/s Statistics – Revisit..

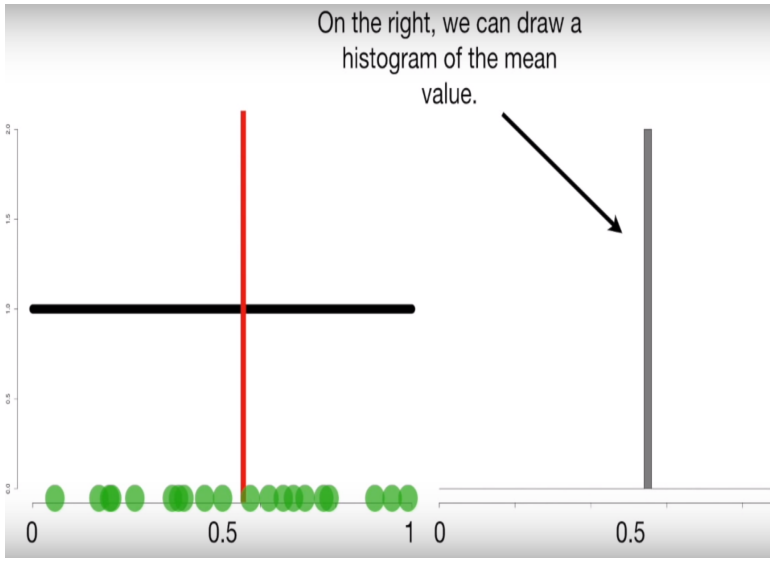


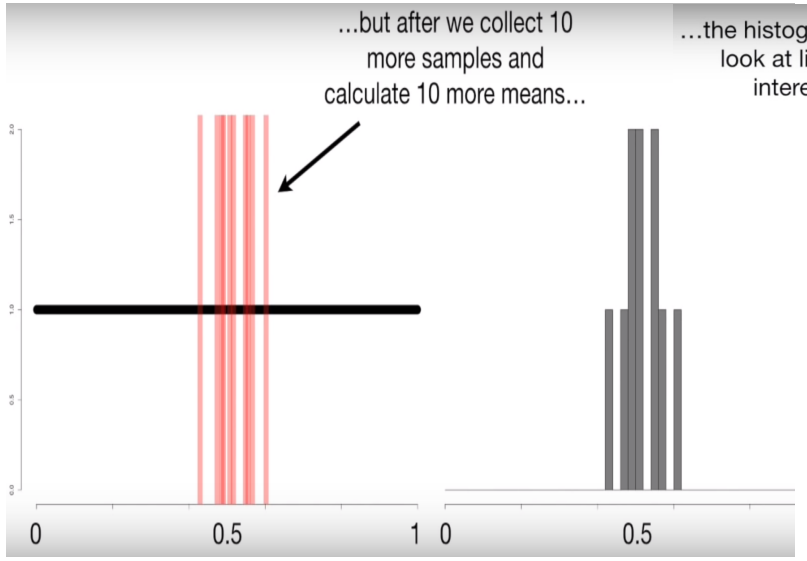
	SAMPLE	POPULATION
MEAN	$\bar{x}$	μ
STANDARD DEVIATION	s	σ
	STATISTIC	PARAMETER

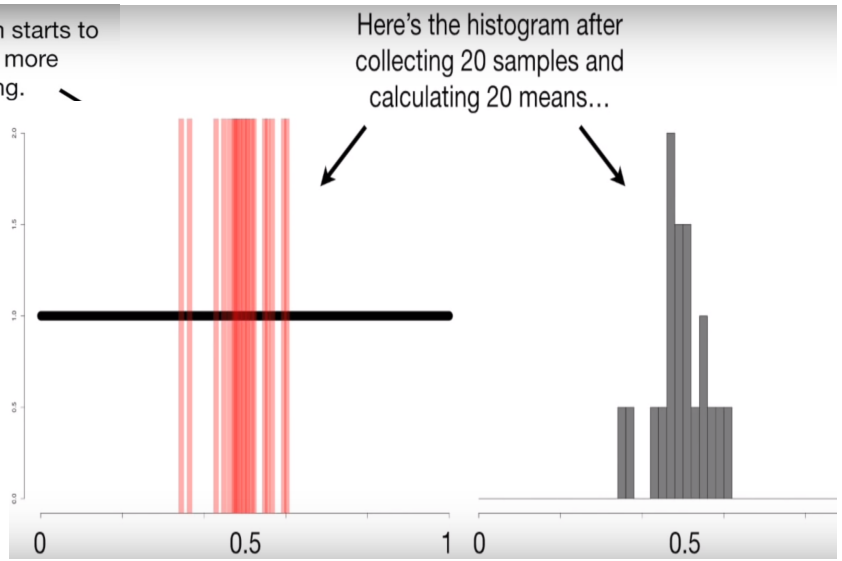
Central Limit Theorem – Where it is applicabl

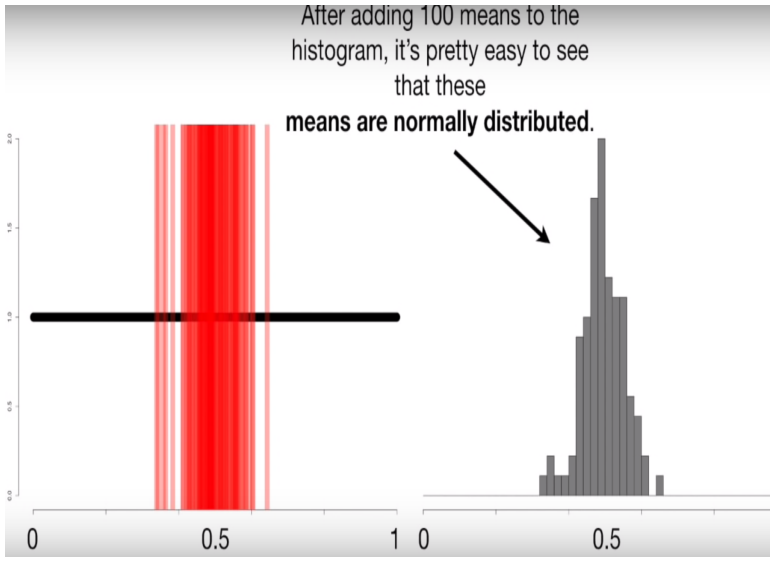




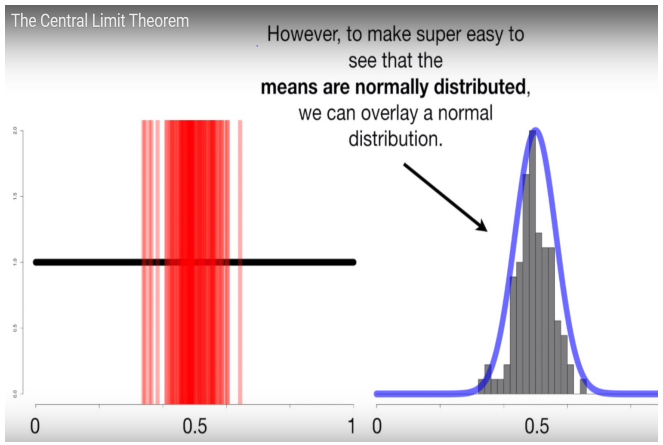


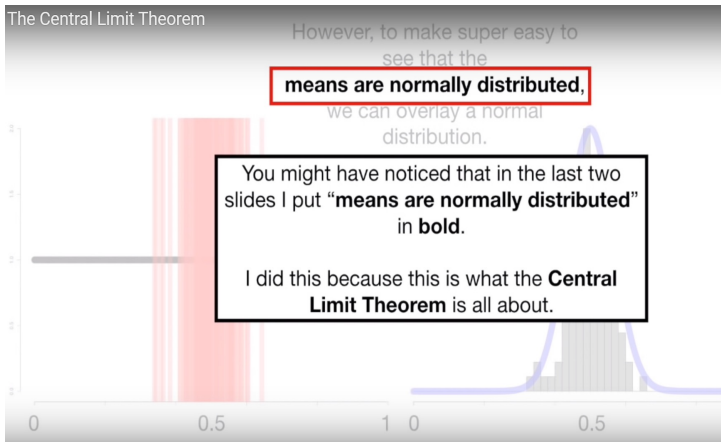






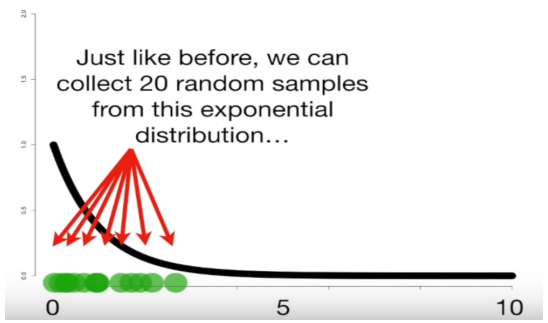
Parameter v/s Statistics – Revisit..

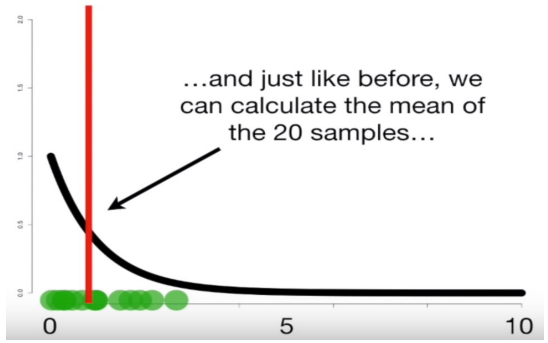


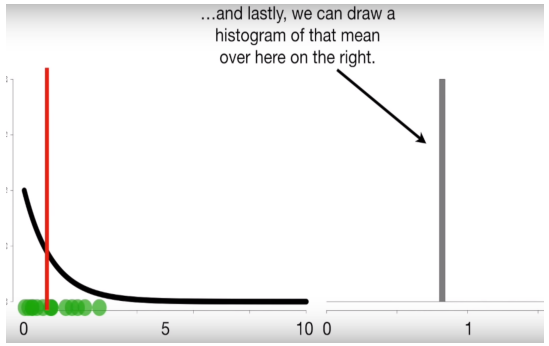


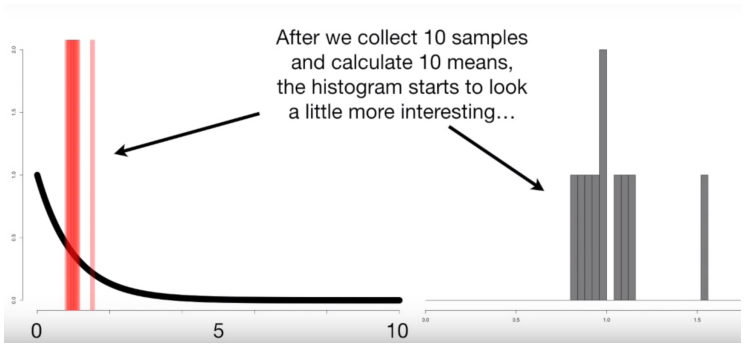
Even though these means were calculated using data from a uniform distribution…

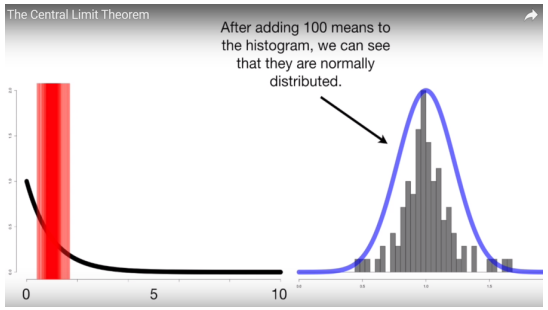
…the means themselves are not uniformly distributed. Instead, the **means are normally distributed**.

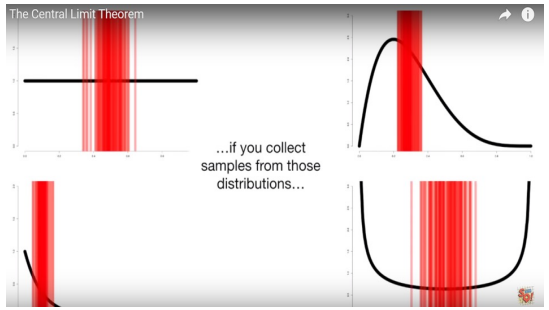


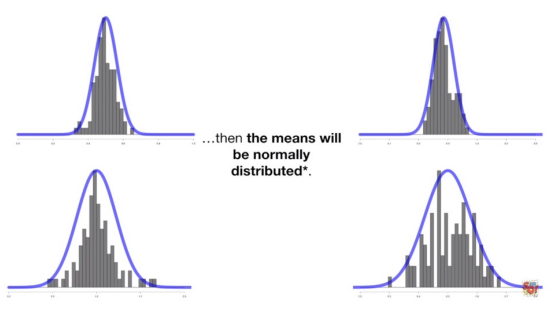












When we do an experiment, we don't always know what distribution our data comes from

To this, **The Central Limit Theorem** says, **"Who Cares???"**

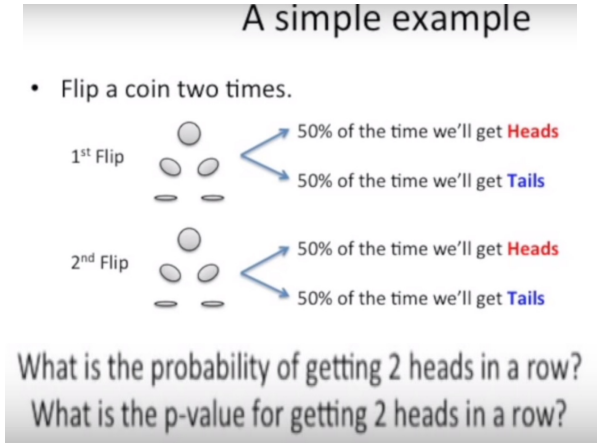
NOTE: Out there in the wild some folks say that in order for the **Central Limit Theorem** to be true, the sample size must be at least **30**.

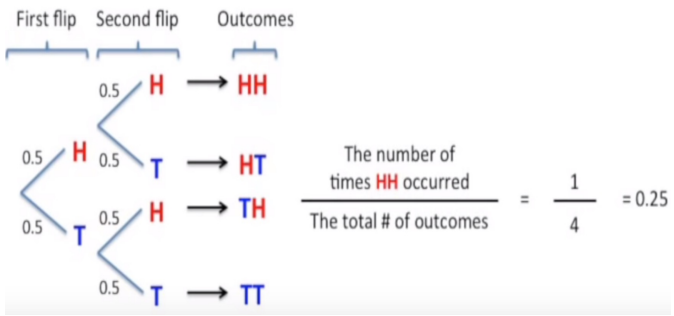
P-values – What are they and where they are coming from?

People often then think that "p-value" means "probability".

They are related, but not the same.

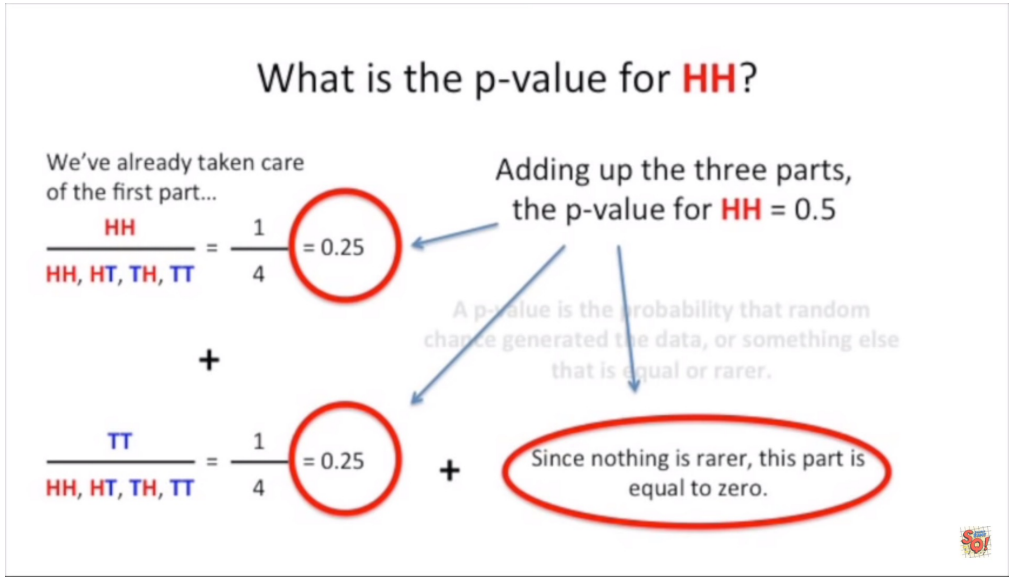
Let's look at a simple example to learn more…





Definition of p-value

A p-value is the probability that random chance generated the data, or something else that is equal or rarer.

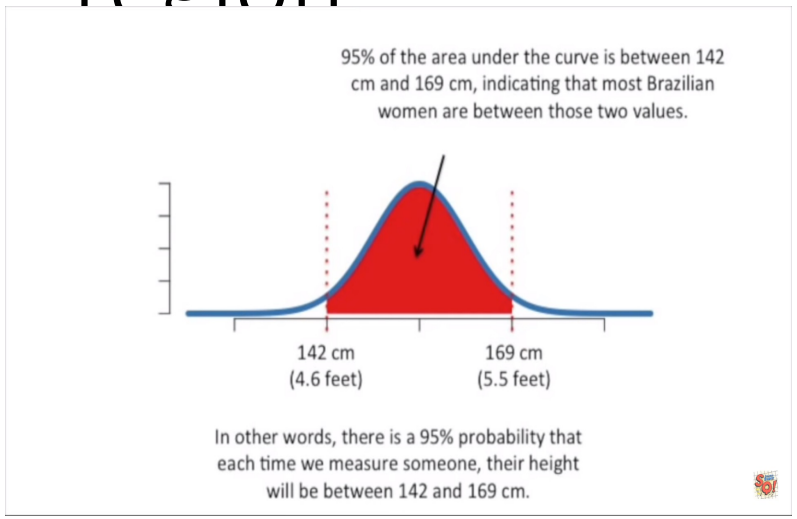


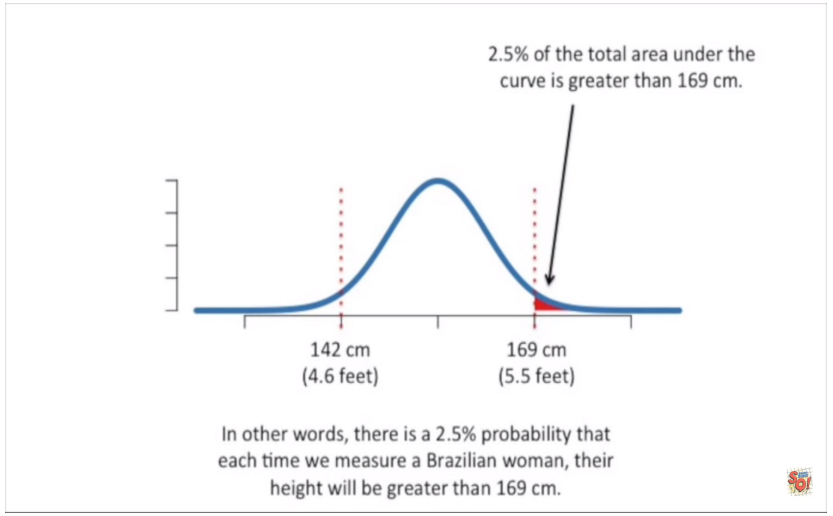
The **probability** of getting HH is **0.25**

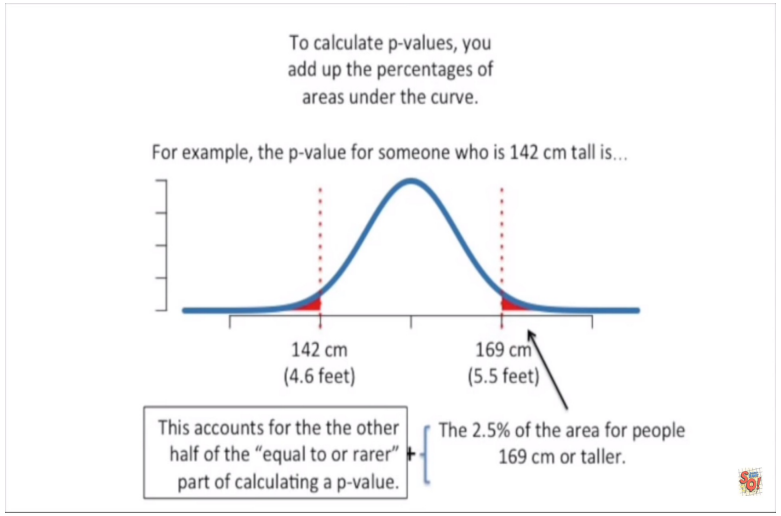
In this case, these are not equal.

The **p-value** for getting HH is **0.5**

P-values – represented with respect to critical region





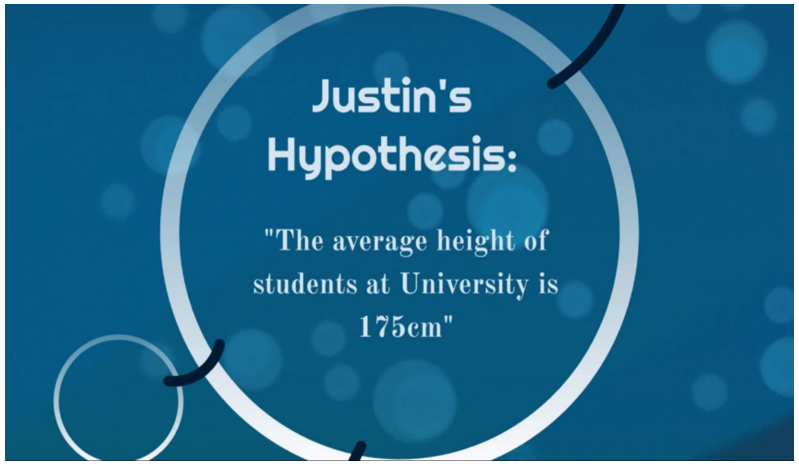


- The **probability** of getting the observed difference or greater when H_o is true.
- A p-value is a measure of how much evidence we have against the null hypothesis.
- Is the lowest significance level such that we will still reject H_o. For a two tailed test, we use twice the table value to find p, and for a one tailed test, we use the table value
- Ranges from 0.0 - 1.0
- If $p \geq 0.05$, then there is no statistical evidence of a difference existing.
- Called **observed** level of significance

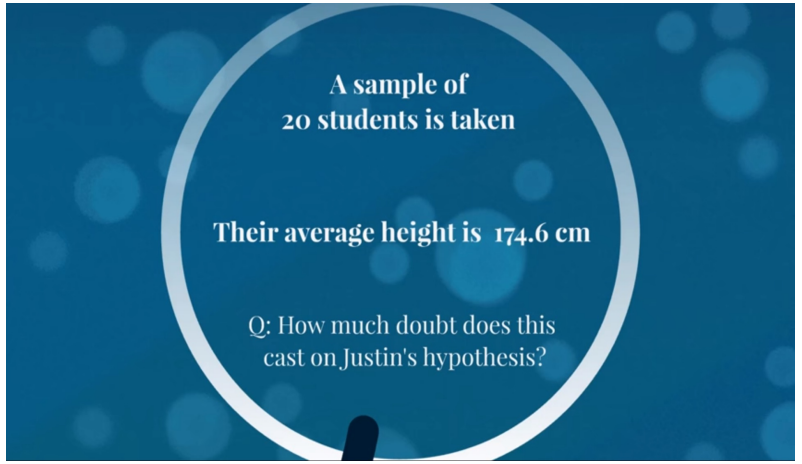
$P < \alpha$: Reject H_o
$P > \alpha$: Accept H_o

Hypothesis Testing – Formulation

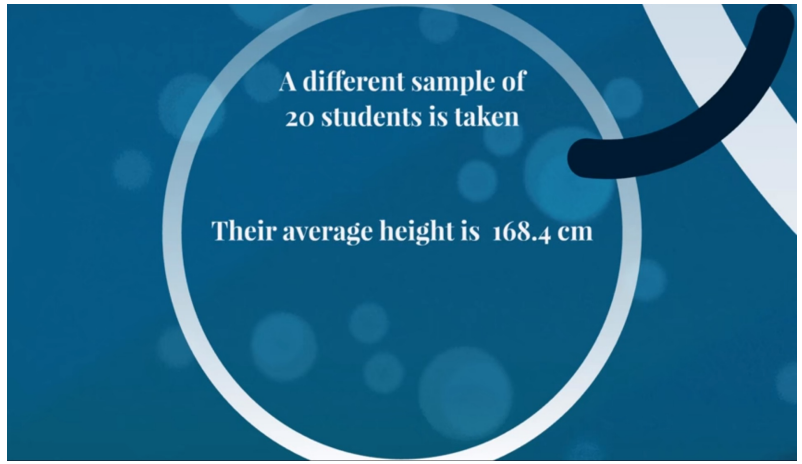
Hypothesis



Sample 1



Sample 2

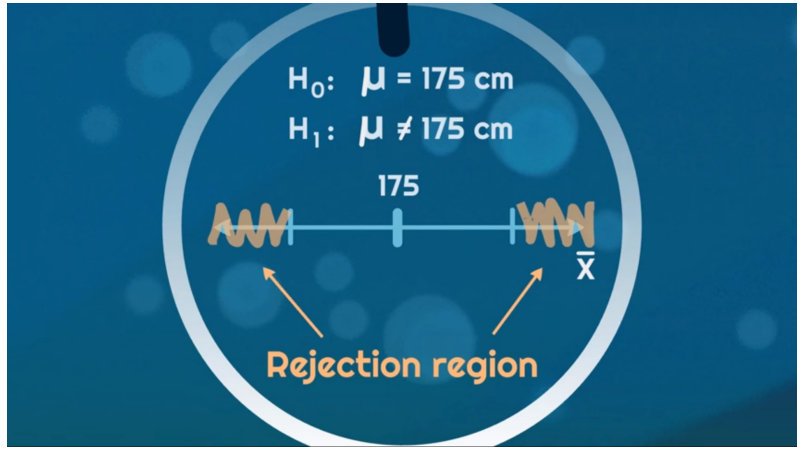


Q- Is 168.4 far enough from 175 (Mean) for us to be able to reject our hypothesis??

Theoretically, it looks like 174.6 (1st Sample) is not very far from 175 but 168.4 (2nd Sample) looks a bit far from 175

<u>To answer this statistically</u> – we will only reject our hypothesis if 168.4 or 174.6 fall into reject region

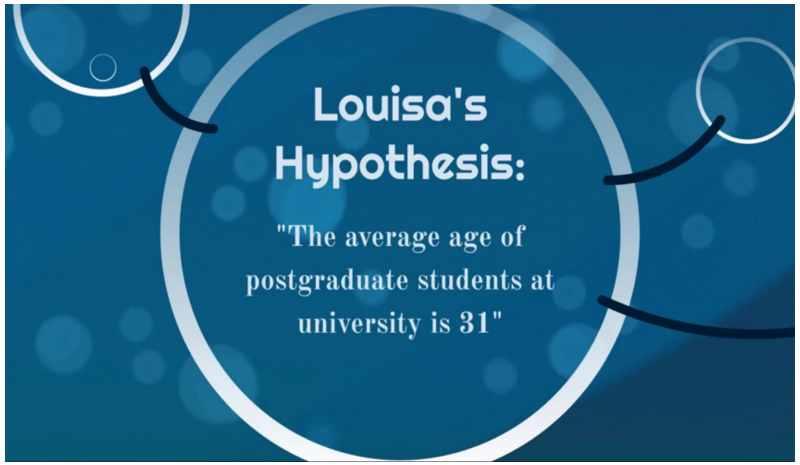
$\overline{X}_2 = 168.4$ $\overline{X}_1 = 174.6$ $\mu = 175$



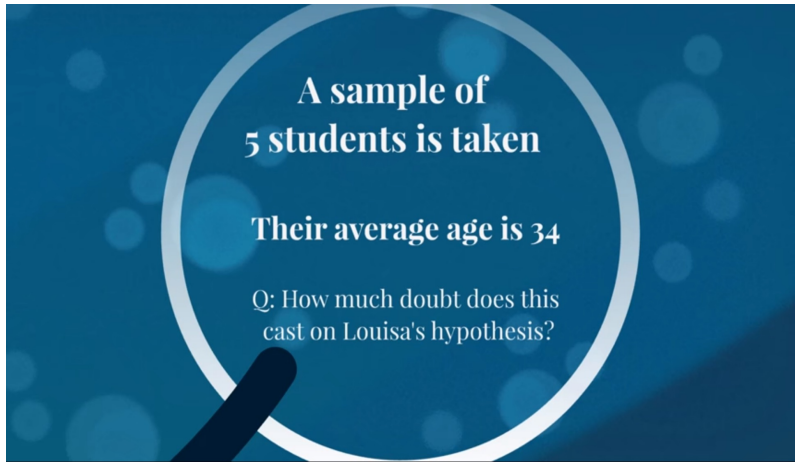
Note - Hypothesis is all about parameters not statistics

Hypothesis Testing – Formulation contd…

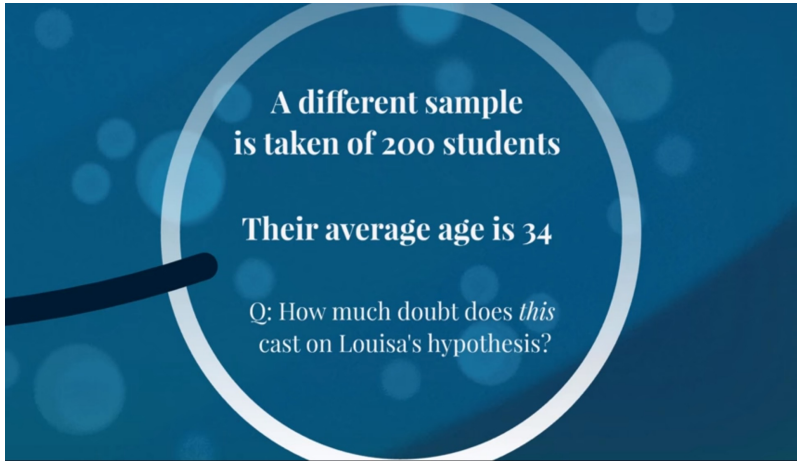
Hypothesis



Sample 1



Sample 2

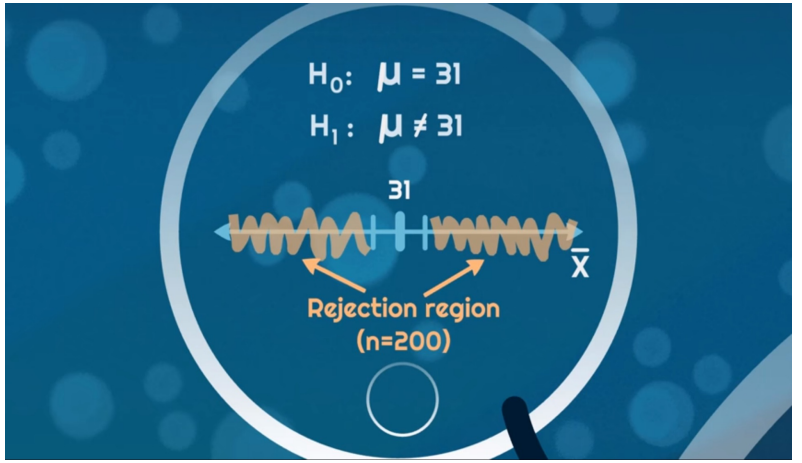


Q- Is 34 far enough or different from 31 (Mean) for us to be able to reject our hypothesis??

Theoretically, it looks 34 (1st Sample) is not different from actual mean(random chance) but 34 (2nd Sample), where sample size is 200 looks different from actual

<u>To answer this statistically</u> – we will only reject our hypothesis if 168.4 or 174.6 fall into reject region

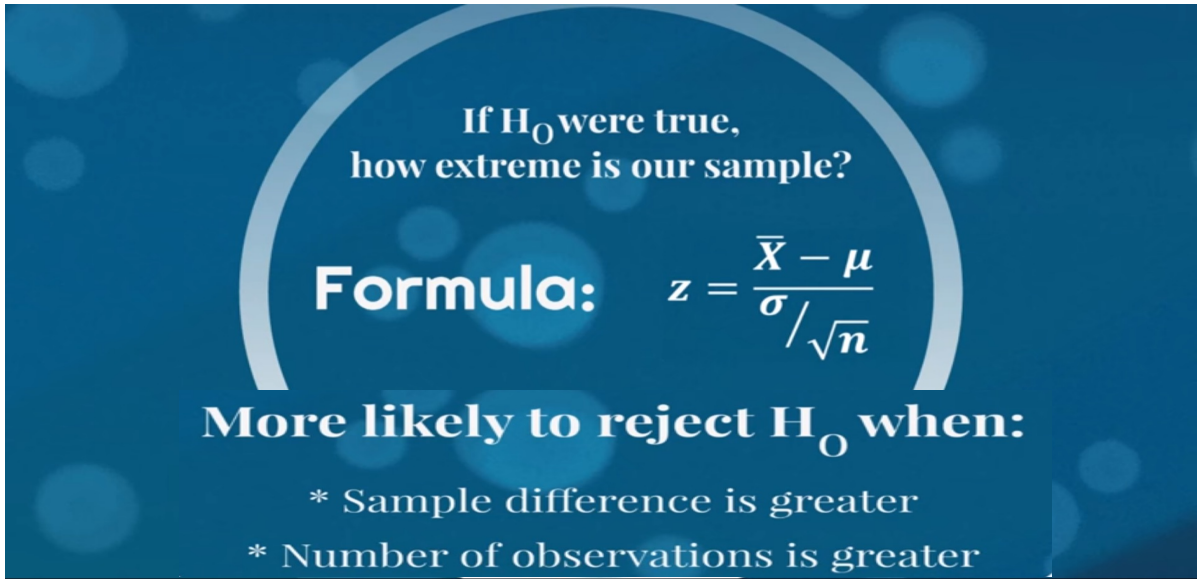
$\overline{X}_2 = 34$ $\overline{X}_1 = 34$ $\mu = 175$

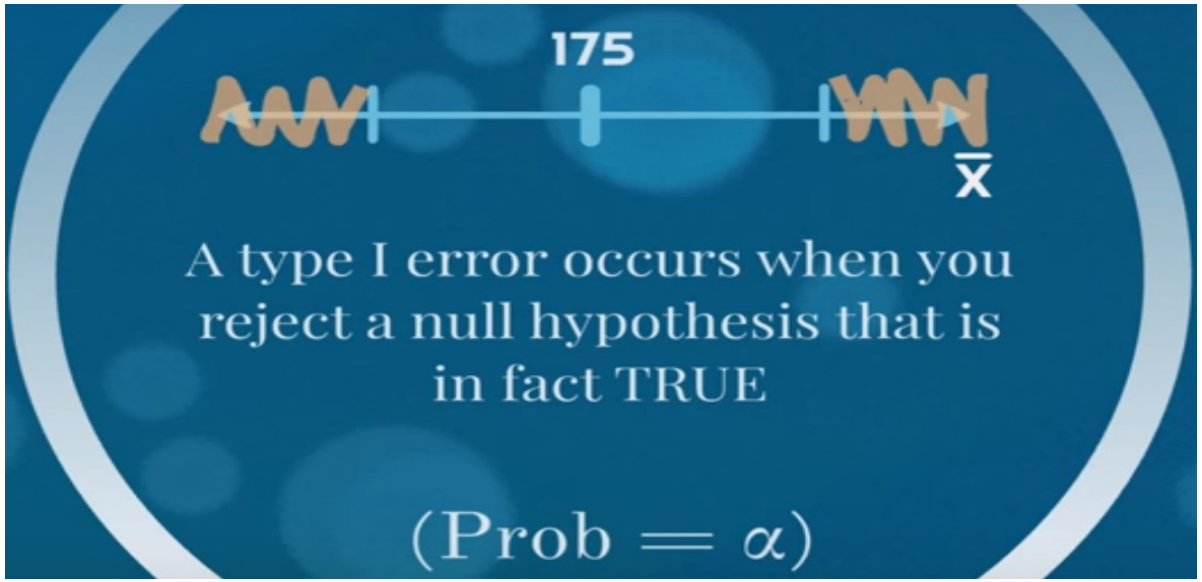


Note – Sample size is drastically different between two samples

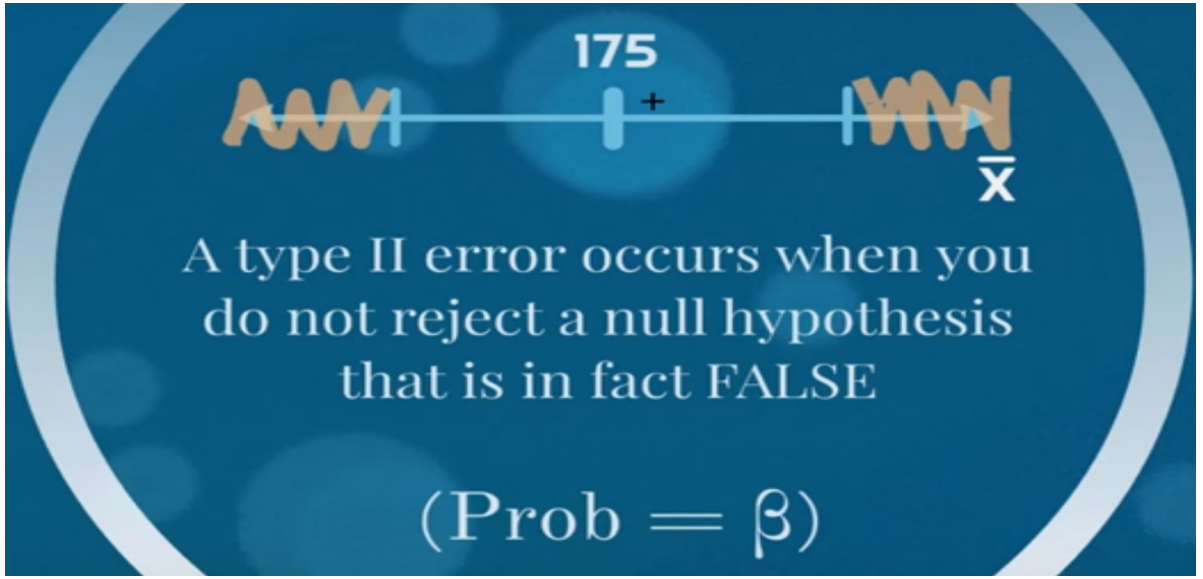
If null hypothesis is true, How extreme is our sample?

Hypothesis Testing – If H_0 is true, how extreme is our sample?





Always a trade-off between type I error and type II error



<u>Power of a test</u> – Probability of rejecting the H_0 (null hypothesis) when it is false

Q - Why we always want to reject the null hypothesis?

Hypothesis Testing for μ when σ is known

Testing for μ when σ is known

Q1: The manager of a department store is thinking about establishing a new billing system for the store's credit customers. She determines that the new system will be cost effective only if the mean monthly account is greater than $70. A random sample of 200 monthly accounts is drawn for which the sample mean account is $74. The manager knows that the accounts are normally distributed with a standard deviation of $30. Is there enough evidence at the 5% level of significance to conclude that the new system will be cost effective?

1

Testing for μ when σ is known

1. STATE NULL AND ALTERNATE HYPOTHESES

H_0: $\mu = 70$

H_1: $\mu > 70$

"....she determines that the new system will be cost effective only if the mean monthly account is greater than $70 ... can we conclude that the new system will be cost effective?"

Note – The alternative hypothesis (H_1) contains values for which we are seeking evidence for

Since we start with a counter here and collect the evidence (against it) null hypothesis, and based on that we reject the null or do not reject the null (never accept the null)

Criminal Trial example – H_0: "the defendant is not guilty"
H_1: "the defendant is guilty"

Hypothesis Testing for μ when σ is known

2 Testing for μ when σ is known

2. CALCULATE TEST STATISTIC

$\bar{X}$	74
σ	30
n	200

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{74 - 70}{30/\sqrt{200}}$$

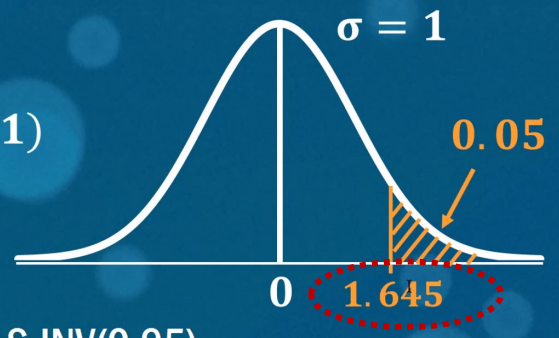
$$z = 1.8856$$

But where is the evidence against the null or 1.8856 is too far from mean?

3 Testing for μ when σ is known

3. CONSIDER DECISION RULE

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



z-crit 1.645 =NORM.S.INV(0.95)

Reject if z > 1.645

5 Testing for μ when σ is known

5. CONCLUSION

There is enough evidence at the 5% level of significance to suggest that the mean monthly account is greater than $70.

4 Testing for μ when σ is known

4. STATE REJECTION DECISION

Reject H0 at 5% level of significance as z > 1.645 (z = 1.8856)

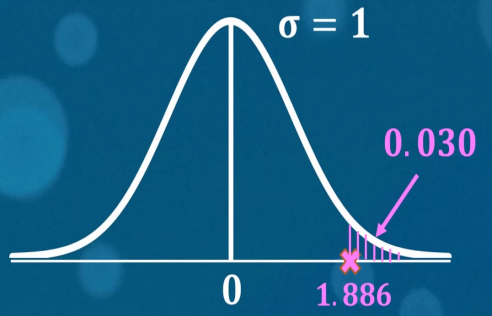
OR

Reject H0 at 5% level of significance as p < 0.05 (p=0.030)

3a Testing for μ when σ is known

[3a. CALCULATE p-VALUE]

The p-value is the probability of getting a sample as extreme as ours, given the null hypothesis is TRUE



p-value =1-NORM.S.DIST(1.886,TRUE) = 0.030

Hypothesis Testing for μ when σ is known

Testing for μ when σ is known

Q2: A new toll road is being built and financed on the expectation that 8,500 cars will use it per day. In the first 30 days of its operation, a daily average of 8,120 cars were found to have used the toll road. Using the 1% level of significance, test whether the expectation was incorrect. (Assume that the distribution of daily road users is normally distributed with a standard devation of 950)

1

Testing for μ when σ is known

1. STATE NULL AND ALTERNATE HYPOTHESES

$H_0: \mu = 8,500$

$H_1: \mu \neq 8,500$

Note – The alternative hypothesis (H_1) contains values for which we are seeking evidence for

Since we start with a counter here and collect the evidence (against it) null hypothesis, and based on that we reject the null or do not reject the null (never accept the null)

Criminal Trial example – H_0: "the defendant is not guilty"
H_1: "the defendant is guilty"

Hypothesis Testing for μ when σ is known

2 Testing for μ when σ is known

2. CALCULATE TEST STATISTIC

$\bar{X}$ 8120
σ 950
n 30

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{8120 - 8500}{950/\sqrt{30}}$$

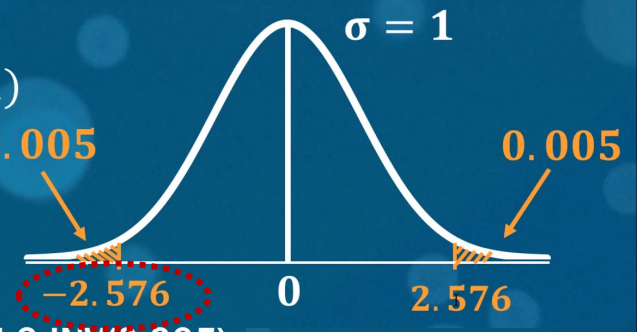
$$z = -2.191$$

But where is the evidence against the null or -2.191 is too far from mean?

3 Testing for μ when σ is known

3. CONSIDER DECISION RULE

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



z-crit 2.576 =NORM.S.INV(0.995) Reject if z>2.576 OR z<-2.576

5 Testing for μ when σ is known

5. CONCLUSION

There is NOT enough evidence at the 1% level of significance to suggest that the daily average of cars using the road is DIFFERENT from 8,500.

4 Testing for μ when σ is known

4. STATE REJECTION DECISION

Do not reject H0 at 1% level of significance as -2.576<z<2.576 (z = -2.191)

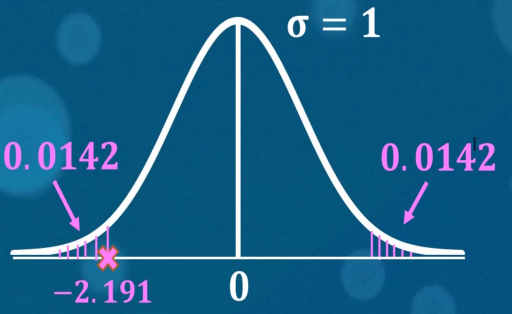
OR

Do not reject H0 at 1% level of significance as p > 0.01 (p = 0.0285)

3a Testing for μ when σ is known

[3a. CALCULATE p-VALUE]

The p-value is the probability of getting a sample as extreme as ours, given the null hypothesis is TRUE



p-value =NORM.S.DIST(-2.191,TRUE)*2 = 0.0285

t Distribution – The problem with small samples

* Solved the problem of "small sample statistics"

Underlying distribution is NORMAL

Population standard deviation unknown

Sample size is too small for C.L.T to apply

* The following measures would be t-distributed

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad \frac{b - \beta}{SE(b)} \qquad \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{\sqrt{n_1}}\right) + \left(\frac{s_2^2}{\sqrt{n_2}}\right)}}$$

If the underlying distribution in normal and standard deviation is known, then sample mean will also follow normal distribution ~

t Distribution

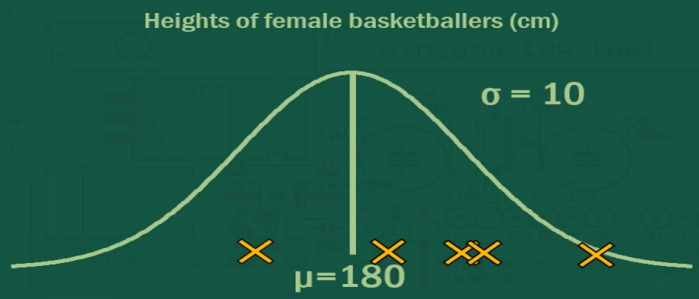
SAMPLING RECAP!

* Take a sample of five observations from a normally distributed population

  [183 , 170 , 189 , 191 , 203]

* Find the average of that sample

  $\bar{X}$ = 187.2 cm



* How would such a sample mean (of size 5) be distributed?

$$\bar{x} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right) \qquad \bar{x} \sim N\left(180, \left(\frac{10}{\sqrt{5}}\right)^2\right)$$

t Distribution – Formulation and Visualization

t Distribution

SAMPLING RECAP!

* Imagine we are TESTING the population mean value of 180cm by using our sample

[183 , 170 , 189 , 191 , 203]

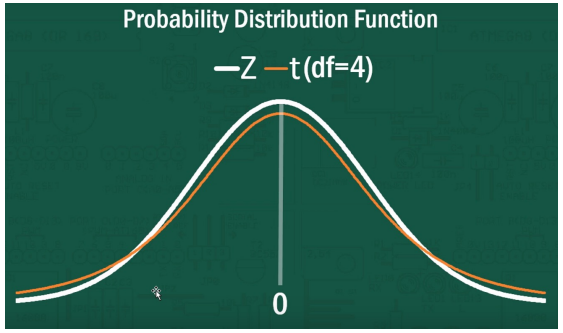
$\bar{X}$ = 187.2 cm

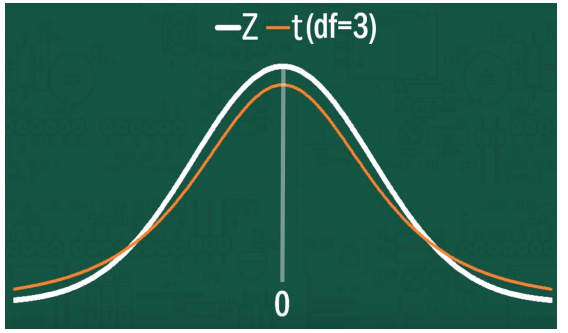
s = 12.05 cm

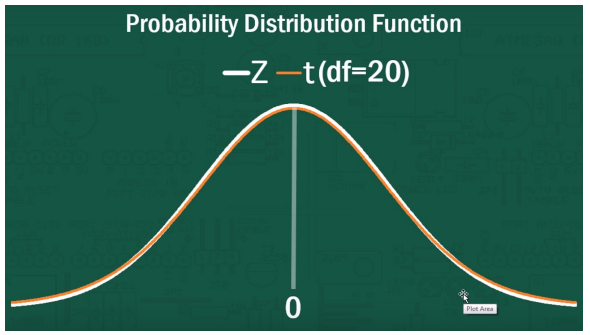
* H_0: μ =180 H_1: μ ≠ 180

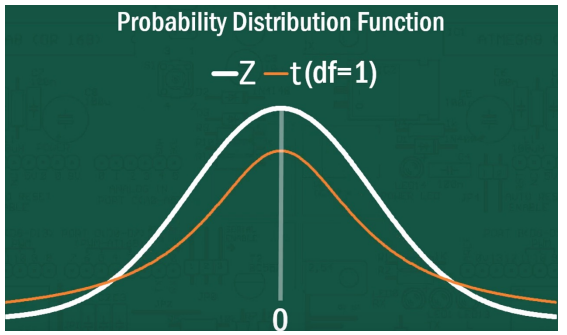
$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim z \longrightarrow \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1} \qquad t = \frac{187.2-180}{12.05/\sqrt{5}} \sim t_4$$

* We need to adjust for the additional uncertainty around s.
* The smaller the sample size, the more uncertain we are.









t Distribution

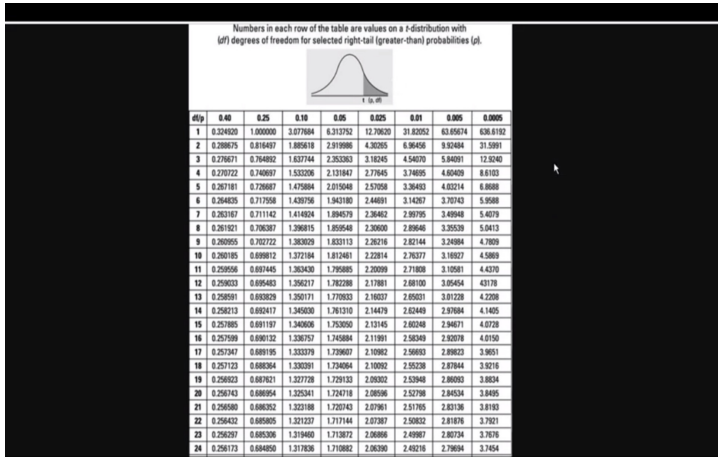
Probability Distribution Function

$$PDF = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\left(1+\frac{x^2}{n-1}\right)^{-\frac{n}{2}}$$

t Distribution

Cumulative Distribution Function

=T.DIST(x,DF,TRUE)



Numbers in each row of the table are values on a t-distribution with (df) degrees of freedom for selected right-tail (greater-than) probabilities (p).

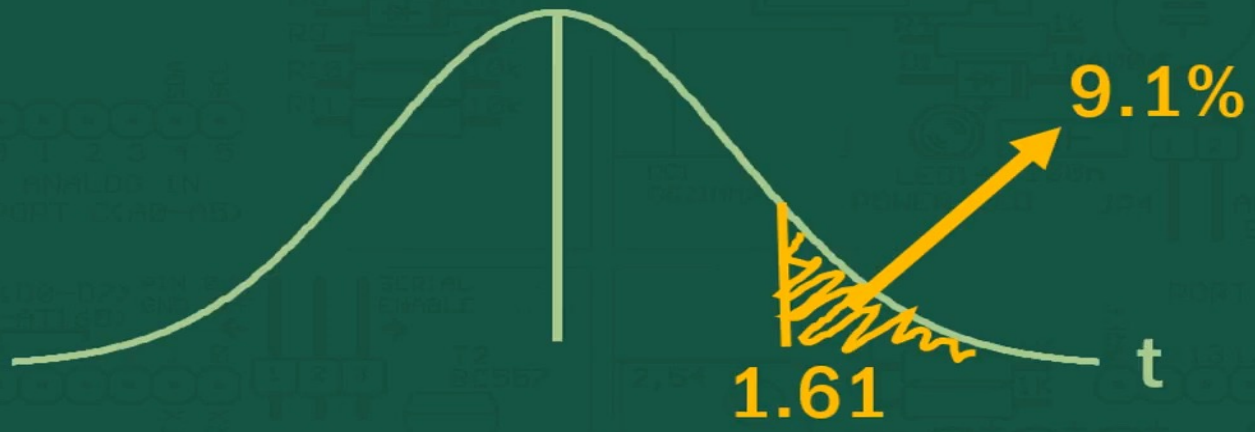
df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454

Practice Question - 1

t Distribution

Cumulative Distribution Function

What proportion of the t-distribution (with 4 df) exists above t=1.61?



=1-T.DIST(1.61,4,TRUE)

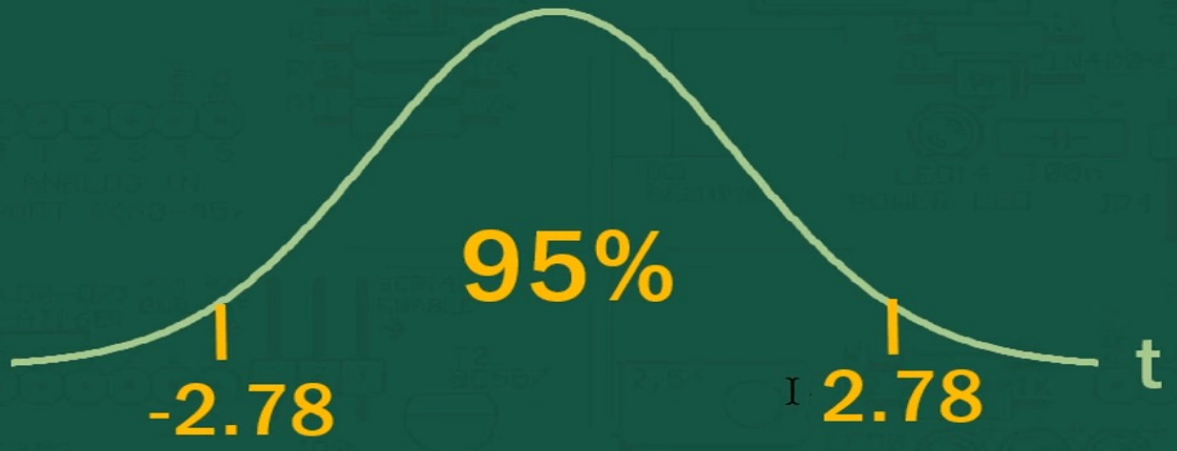
=0.091

Practice Question - 2

t Distribution

Cumulative Distribution Function

What t statistic (with 4 df) provides 2.5% in the upper tail?



=T.INV(0.975,4)

=2.78

Hypothesis Testing for μ - When σ is unknown

Testing for μ when σ is unknown

An online fashion store called Showdonkey advertises that its average delivery time is less than six hours for local deliveries. A random sample of the amount of time taken to deliver packages to an address in Stanmore produced the following delivery times (rounded to the nearest hour):

7 3 4 6 10 5 6 4 3 8

Is there sufficient evidence to support Showdonkey's advertisement, at the 5% level of significance?

1

Testing for μ when σ is unknown

1. STATE NULL AND ALTERNATE HYPOTHESES

$$H_0: \mu = 6$$

$$H_1: \mu < 6$$

"....Is there sufficient evidence to support Showdonkey's advertisement"

Note – The alternative hypothesis (H_1) contains values for which we are seeking evidence for

Since we start with a counter here and collect the evidence (against it) null hypothesis, and based on that we reject the null or do not reject the null (never accept the null)

Criminal Trial example – H_0: "the defendant is not guilty"
H_1: "the defendant is guilty"

Hypothesis Testing for μ - When σ is unknown

2a

2. CALCULATE TEST STATISTIC

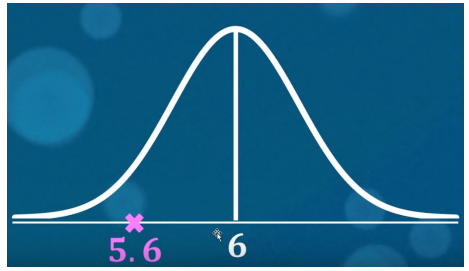
7	3	4	6	10	5	6	4	3	8

$\bar{x}$ 5.6 =AVERAGE(B6:K6)

s 2.27 =STDEV.S(B6:K6)

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

But where is the evidence against the null -0.557 is too far from mean?



2

Testing for μ when σ is unknown

2. CALCULATE TEST STATISTIC

$\bar{x}$	5.6
s	2.3
n	10

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{5.6 - 6}{2.3/\sqrt{10}}$$

$$t = -0.557$$

4

Testing for μ when σ is unknown

4. STATE REJECTION DECISION

Do not reject H0 at 5% level of significance as t > -1.833 (t = -0.557)

OR

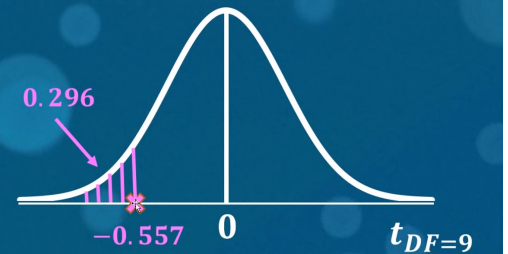
Do not reject H0 at 5% level of significance as p > 0.05 (p=0.296)

3a

Testing for μ when σ is unknown

[3a. CALCULATE p-VALUE]

The p-value is the probability of getting a sample as extreme as ours, given the null hypothesis is TRUE



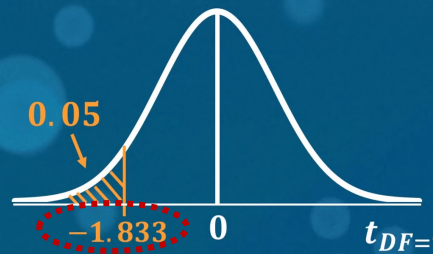
p-value =T.DIST(-0.557,9,TRUE) = 0.296

3

Testing for μ when σ is unknown

3. CONSIDER DECISION RULE

$$t = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim t_{DF}$$



t-crit -1.833 =T.INV(0.05,9

Reject if t < -1.833

5

5. CONCLUSION

While the sample mean was 5.6 hours, there is NOT enough evidence at the 5% level of significance to infer that the population average time for delivery is less than 6 hours.

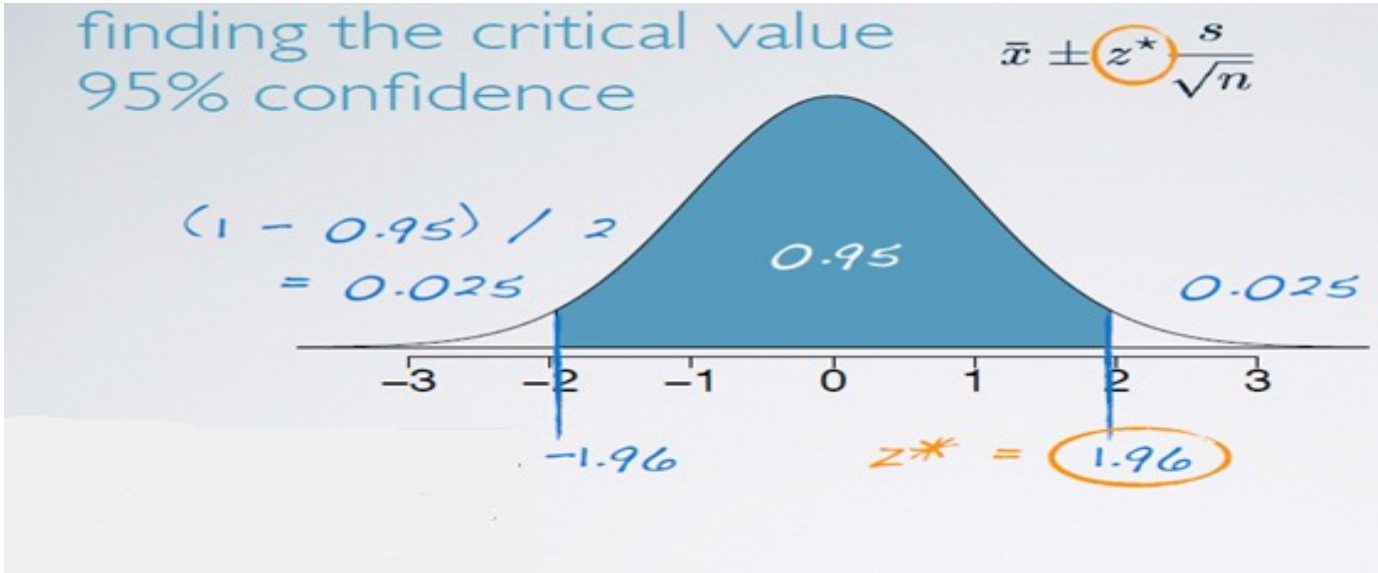
Confidence Interval – Introduction

Confidence interval for a population mean: Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).

$$\bar{x} \pm z^\star \frac{s}{\sqrt{n}}$$

Conditions for this confidence interval:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** n ≥ 30, larger if the population distribution is very skewed.



Second decimal place					
0.07	0.06	0.05	0.04	0.00	z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

Confidence Interval – Contd..

A sample of 50 college students were asked how many exclusive relationships they've been in so far. The students in the sample had an average of 3.2 exclusive relationships, with a standard deviation of 1.74. In addition, the sample distribution was only slightly skewed to the right. Estimate the true average number of exclusive relationships based on this sample using a 95% confidence interval.

1. random sample & 50 < 10% of all college students

We can assume that the number of exclusive relationships one student in the sample has been in is independent of another.

2. $n > 30$ & not so skewed sample

We can assume that the sampling distribution of average number of exclusive relationships from samples of size 50 will be nearly normal.

$n = 50$

$\bar{x} = 3.2$

$s = 1.74$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.246$$

$$\bar{x} \pm z^* SE = 3.2 \pm 1.96\,(0.246)$$

$$= 3.2 \pm 0.48$$

$$= (2.72, 3.68)$$

We are 95% confident that college students on average have been in 2.72 to 3.68 exclusive relationships.

Summary of Testing Tables

	Jury Trial			Hypothesis Test	
	Actual Situation			Actual Situation	
Verdict	Innocent	Guilty	Decision	H_0 True	H_0 False
Innocent	Correct	Error (II)	Do not reject H_0	$1-\alpha$	Type II error (β)
Guilty	Error (I)	Correct	Reject H_0	Type I error (α)	Power ($1-\beta$)

Type I Error

- Rejecting True Null Hypothesis ("*False Positive*")
- Probability of Type I error is α
  - Called Level of Significance
  - Set by researcher; provides critical values for the test
  - Called **Rejection Region** of Sampling Distribution / Typical values are 0.01, 0.05, 0.10

Type II Error

- Do not reject False Null Hypothesis ("*False Negative*")
- Probability of Type II error is β

Z Test

Perform z test when:

- ❑Random sample follows a Normal distribution
- ❑ Mean is known
- ❑ Population standard deviation is known
- ❑ Large sample (n>= 30)

Z test can be used

- ❑ For testing mean of single sample
- ❑ For testing means of two independent samples

Z Tests - Example

Maxwell's Hot Chocolate is concerned about the effect of the recent years long coffee advertising campaign on hot chocolate sales. The average hot chocolate sales two years ago was 984.7 pounds and the standard deviation was 99.6 pounds.

Maxwell randomly selected 30 weeks from the past year and found average sales to be 967.1 pounds.

Test:
a. Whether the hot chocolate sales have decreased?
b. At 5% level of significance help the Maxwell hot chocolate to take a decision?

Ho: $\mu = 984.7$ vs. H1: $\mu < 984.7$

Under Ho, the test statistics is:

$$Z = \frac{(\bar{x} - \mu_o)}{\sigma / \sqrt{n}}$$

Numerator : 967.1 – 984.7 = -17.6
Denominator: 99.6 / SQRT(30)
Z = -0.96 p value = <0.0001

What do we conclude?

The result is significant at $p < .05$:
Reject the Null Hypothesis

Note: In Excel Use NORMSDIST function to compute the p value.

T Test

The t test tells you how significant the differences between groups are; In other words it lets you know if those differences (measured in means/averages) could have happened by chance.

Perform t test when:

- ❑ Random sample follows a Normal distribution
- ❑ Mean is known
- ❑ Population Variance is unknown
- ❑ Small sample (n< =30)

T test can be used

- ❑ For testing mean of single sample
- ❑ For testing means of two independent samples
- ❑ Difference between means of dependent sample

T Tests - Example

You seek to determine whether differences exist in monthly sales between the new package design and the old package design of a laundry stain remover. The new package was test marketed over a period of one month in a sample of supermarkets in a particular city.

A random sample of ten pairs of supermarkets was matched according to weekly sales volume and a set of demographic characteristics. The data collected for this study are as follows:

Subject #	Score 1	Score 2
1	3	20
2	3	13
3	3	13
4	12	20
5	15	29
6	16	32
7	17	23
8	19	20
9	23	25
10	24	15
11	32	30

Subject #	Score 1	Score 2	X-Y
1	3	20	-17
2	3	13	-10
3	3	13	-10
4	12	20	-8
5	15	29	-14
6	16	32	-16
7	17	23	-6
8	19	20	-1
9	23	25	-2
10	24	15	9
11	32	30	2

New Package = Score1
Old Package = Score
Subject: Supermarket

T Tests – Example Cont.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**

$$t = \frac{(\Sigma D)/N}{\sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{N}}{(N-1)(N)}}}$$

ΣD: Sum of the differences (Sum of X-Y from Step 2)
ΣD^2: Sum of the squared differences (from Step 4)
$(\Sigma D)^2$: Sum of the differences (from Step 2), squared.

$$t = \frac{(\Sigma D)/N}{\sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{N}}{(N-1)(N)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \frac{(-73)^2}{11}}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \frac{5329}{11}}{110}}}$$

$$t = -2.74$$

Step 6: Subtract 1 from the sample size to get the degrees of freedom. We have 11 items, so 11-1 = 10.

Step 7: Find the p-value in the t-table, using the degrees of freedom in Step 6. If you don't have a specified alpha level, use 0.05 (5%). For this sample problem, with df=10, the t-value is 2.228.

Step 8: Compare your t-table value from Step 7 (2.228) to your calculated t-value (-2.74). The calculated t-value is greater than the table value at an alpha level of .05. The p-value is less than the alpha level: p <.05. We can reject the null hypothesis that there is no difference between means.

Note: In Excel, use the TTEST function to compute the p value.

F Test

Perform F test when:

- ❑ Populations from which the random samples are drawn are normal and independent
- ❑ Let $s_1^2$and s_2^2 be the sample variances

F test can be used

- ❑ to test the hypothesis of equality of two population variances
- ❑ to test for equality of several means is carried out by the technique named ANOVA
- ❑ to test the significance of the regression model

F Test: Example

Two chemical companies supply a raw material. The concentration of a particular element in this material is important. The mean concentration for both suppliers is the same, but we suspect that the variability in concentration may differ between the two companies.

The standard deviation of
concentration in a random sample of n1=10 batches produced
by company 1 is s1 = 4.7 grams per liter, while for company 2, a random sample of n2 = 16 batches yields s2= 5.8 grams per liter.

Is there sufficient evidence to conclude that the two population variances differ?

Ho : $\sigma 1^2 = \sigma 2^2$ vs.

Ha : $\sigma 1^2 \neq \sigma 2^2$

Under Ho, the test statistics

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

F = 0.6577
df = 9, 15
p value = 0.733

What should we infer?

Since p value >0.05, we do no have enough evidence to reject the hypothesis.

F Test: Example Cont.

A fuel-economy study was conducted for two German automobiles, Mercedes and Volkswagen. One vehicle of each brand was selected, and the mileage performance was observed for 10 tanks of fuel in each car.

Is there evidence to support the claim that the variability in mileage performance is lesser for a Mercedes than for a Volkswagen?

How should we frame the hypothesis?
Ho : $\sigma 1^2 <= \sigma 2^2$ vs.
Ha : $\sigma 1^2 > \sigma 2^2$
Under the null hypothesis, the test statistics is:

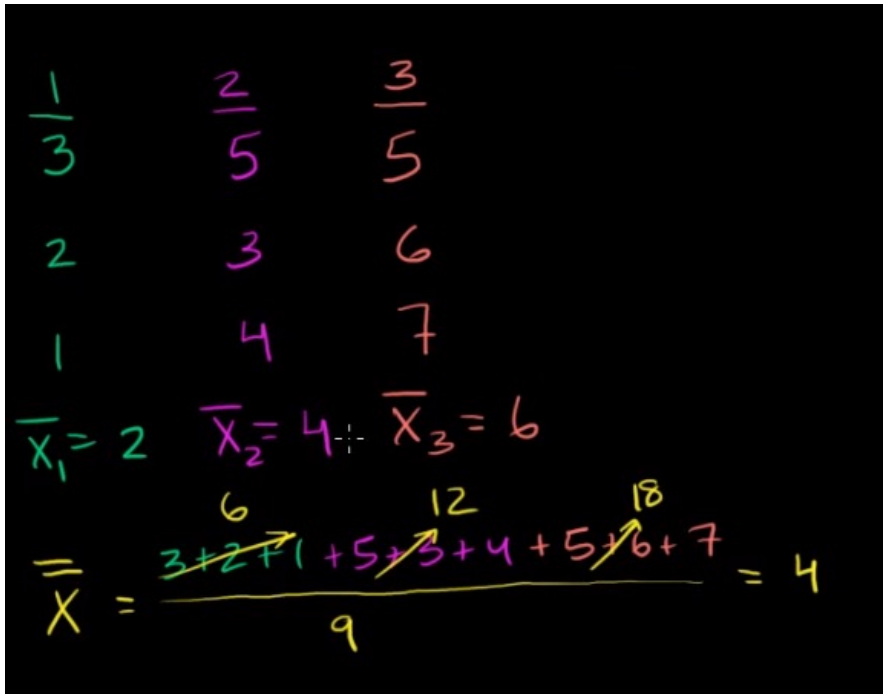
$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Mercedes	Volkswagen
24.7	41.7
24.8	42.3
24.9	41.6
24.7	39.5
24.5	41.9
24.9	42.8
24.6	42.4
24.6	39.9
24.9	40.8
24.8	39.6

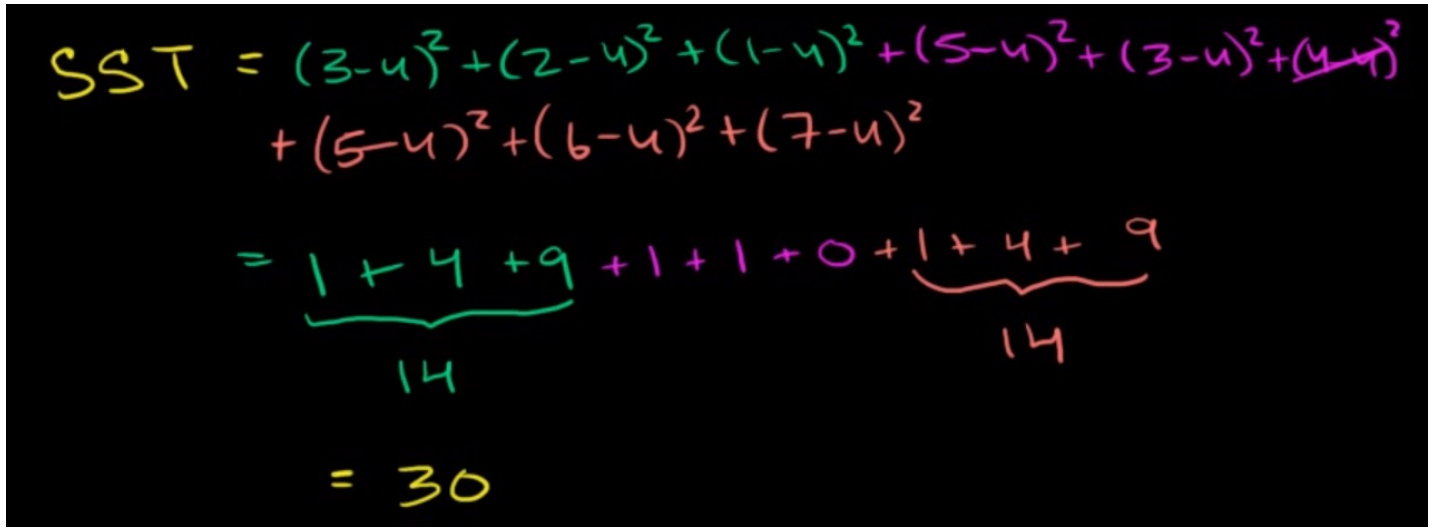
F =0.0137 with df = 9,9
p value = 2.50E-07

What do we infer?

ANOVA Test: Build Up SST



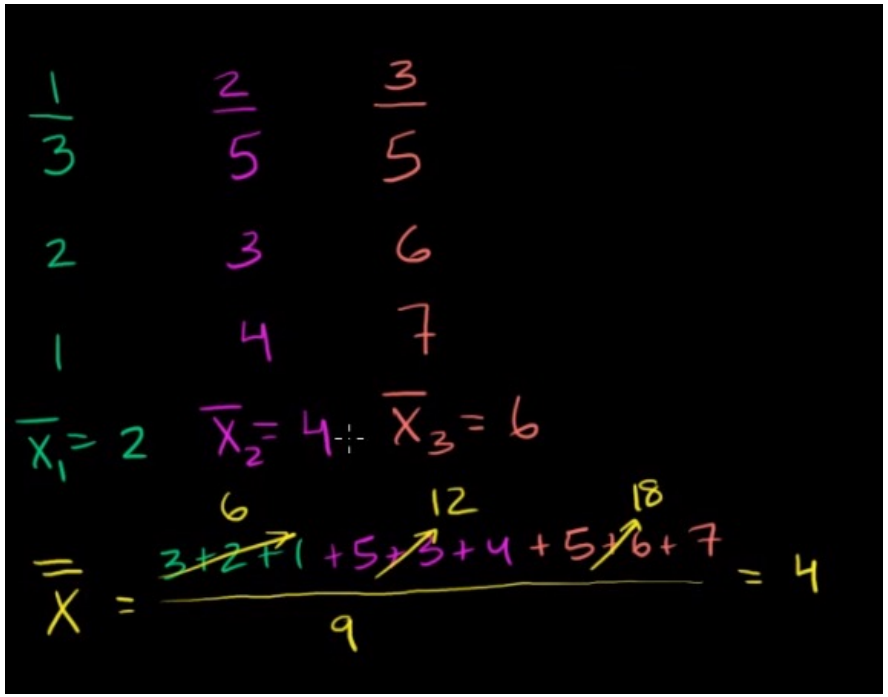
Grand Mean = 4



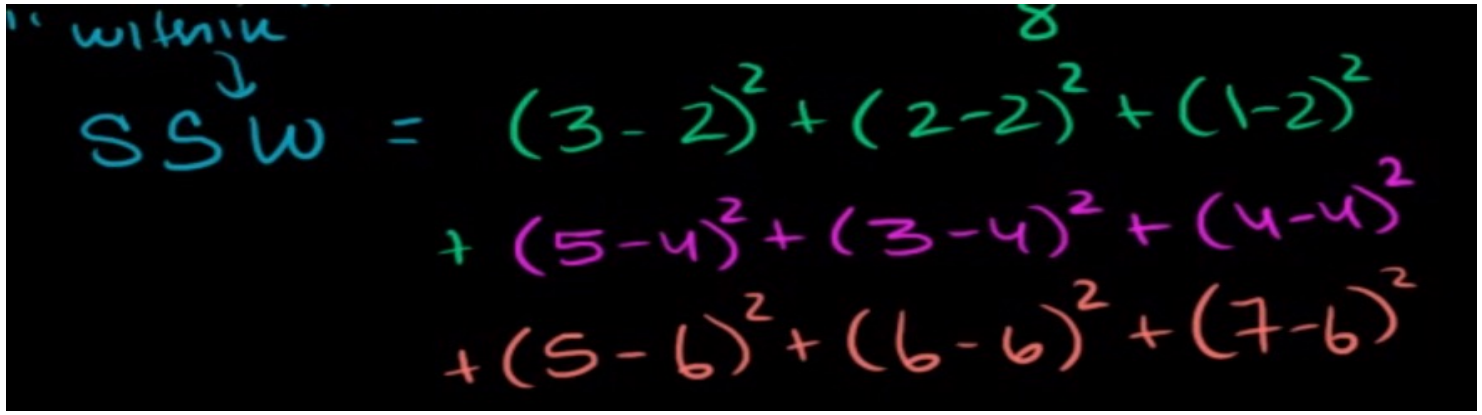
Total Sum of Square: 30
Degrees of Freedom: (m * n) – 1 = 8

NULL Hypothesis: All the group means are equal
ALT Hypothesis: At least one is not Equal

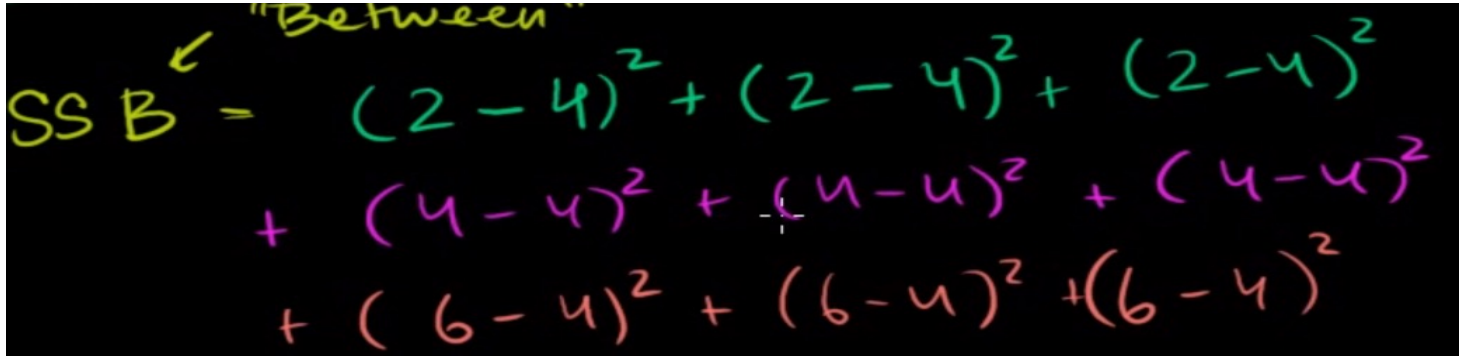
ANOVA Test: Build Up SSW and SSB



Grand Mean = 4

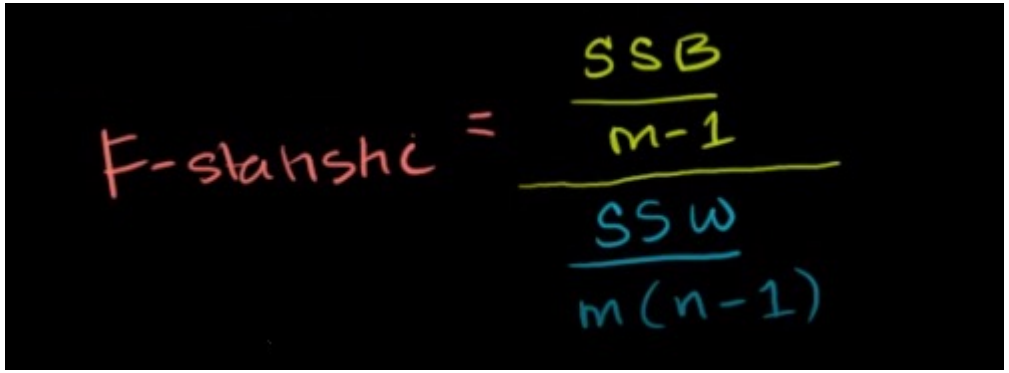


Within Sum of Square: 6
Degrees of Freedom: (n-1) * # of Groups = 3 * 2 = 6



Between Sum of Square: 24
Degrees of Freedom: (n-1) * # of Groups = m-1 = 2

ANOVA Test: F Statistic

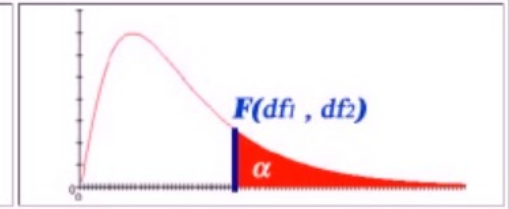


F Value = (SSB / DF of SSB) / (SSW / DF of SSW)
= (24/2) / (6/6) = 12

F Score = 3.46330

Reject the NULL Hypothesis

F Table for α = 0.10



\	df_1=1	2	3	4	5	6	7	8	9	10	12
df_2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804	2.50

Chi Square Tests

Categorical data lead to counts within each category

Tests the association/ independence between nominal(categorical) variables

Null Hypothesis: The 2 variables are independent

Tests the homogeneity between categorical variables

Measure of goodness-of -fit

Its really just a comparison between expected frequencies and observed frequencies among the cells in a cross tabulation table

Requirements for Chi-Square test

Must be a random sample from population

Data must be in raw frequencies

Variables must be independent

Categories for each cell must be mutually exclusive and exhaustive

Using the Chi-Square Test

Often used with contingency tables (i.e., cross tabulations)

- E.g., gender x race

Basically, the chi-square test of independence tests whether the columns are contingent on the rows in the table.

- In this case, the null hypothesis is that there is no relationship between row and column frequencies.

Example

Fast-food chains are evaluated on many variables and the results are summarized periodically in QSR Magazine. One important variable is the accuracy of the order.

You seek to determine, with a level of significance of $\alpha = 0.05$, whether a difference exists in the proportions of food orders filled correctly at Burger King, Wendy's, and McDonald's.

How will you frame the hypothesis?

Ho: No difference exists in the proportion of correct orders among Burger King, Wendy's, and McDonald's.

H1: A difference exists in the proportion of correct orders among Burger King, Wendy's, and McDonald's

Under Ho the test statistics is as follows:

$$\sum \frac{(Oi - Ei)^2}{Ei} \sim \chi^2(m-1, n-1)$$

	Expected Frequencies			
		Burger King	Wendy's	McDonald's
Order Filled Correctly	Yes	431	431	431
	No	69	69	69
	Total	500	500	500

	Fast Food Chain				
		Burger King	Wendy's	McDonald's	Total
Order Filled Correctly	Yes	440	430	422	1292
	No	60	70	78	208
	Total	500	500	500	1500

Example2

The χ^2 Statistic for nominal data

Presume you observe 100 people to see who deposits garbage in the can and who litters. You want to see if there is a difference based on gender.

A person can fall in one of four categories:
- Male, deposits garbage
- Male, litters
- Female, deposits garbage
- Female, litters

	Deposit	Litter	
Females	18	7	25
Males	42	33	75
	60	40	100

To answer this question, you have to figure out what numbers you might expect if everything were left to chance; if H_0 were true—that there is no difference based on gender.

	Deposit	Litter	
Females	18 15	7 10	25
Males	42 45	33 30	75
	60	40	100

Working in a similar method, you can fill in all the expected values.

The further the observed values are from the expected values, the more likely that there really *is* a significant difference.

$$\sum \frac{(Oi-Ei)^2}{Ei} \sim \chi^2(m-1,n-1)$$

	Deposit	Litter	
Females	18 15	7 10	25
Males	42 45	33 30	75
	60	40	100

In this case, that works out to:

$$\frac{(18-15)^2}{15}+\frac{(7-10)^2}{10}+\frac{(42-45)^2}{45}+\frac{(33-30)^2}{30}$$
$$=\frac{9}{15}+\frac{9}{10}+\frac{9}{45}+\frac{9}{30}$$
$$=.6+.9+.2+.3$$
$$=2.0$$

Looking up the value 2.0 in the χ^2 table for 1 degree of freedom, you find the probability of this result is 0.16, so you retain H_0; there's no significant difference based on gender.

Summary of Parametric Test

Test		Null hypothesis	Alternative hypothesis	Test Statistic	Rejection Rule	*p*-value*
One-Sample Test for the Mean		$H_0 : \mu = \mu_0$	$H_1 : \mu \neq \mu_0$	$t_c = \frac{\bar{x} - \mu_0}{(s/\sqrt{n})}$	$\lvert t_c \rvert > t_{\alpha/2,n-1}$	$2 \cdot tcdf(\lvert t_c \rvert, 999, n-1)$
			$H_1 : \mu > \mu_0$		$t_c > t_{\alpha,n-1}$	$tcdf(\lvert t_c \rvert, 999, n-1)$
			$H_1 : \mu < \mu_0$		$t_c < -t_{\alpha,n-1}$	
Difference of Two Means (Independent Samples)	$\sigma_1 \neq \sigma_2$	$H_0 : \mu_1 = \mu_2$	$H_1 : \mu_1 \neq \mu_2$	$z_c = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{SE_1^2 + SE_2^2}}$ $SE_1^2 = S_1^2/n_1 ; SE_2^2 = S_2^2/n_2$	$\lvert z_c \rvert > z_{\alpha/2}$	$2 \cdot normalcdf(\lvert z_c \rvert, 999)$
			$H_1 : \mu_1 > \mu_2$		$z_c > z_\alpha$	$normalcdf(\lvert z_c \rvert, 999)$
			$H_1 : \mu_1 < \mu_2$		$z_c < -z_\alpha$	
	$\sigma_1 = \sigma_2$	$H_0 : \mu_1 = \mu_2$	$H_1 : \mu_1 \neq \mu_2$	$t_c = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{SE_1^2 + SE_2^2}}$ $SE_1^2 = S_p^2/n_1 ; SE_2^2 = S_p^2/n_2$ $S_p^2 = [(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)$	$\lvert t_c \rvert > t_{\alpha/2,n_1+n_2-2}$	$2 \cdot tcdf(\lvert t_c \rvert, 999, n_1+n_2-2)$
			$H_1 : \mu_1 > \mu_2$		$t_c > t_{\alpha,n_1+n_2-2}$	$tcdf(\lvert t_c \rvert, 999, n_1+n_2-2)$
			$H_1 : \mu_1 < \mu_2$		$t_c < -t_{\alpha,n_1+n_2-2}$	
Paired Differences		$H_0 : \mu_D = 0$	$H_1 : \mu_D \neq 0$	$t_c = \frac{\overline{diff}}{(s_{diff}/\sqrt{n})}$	$\lvert t_c \rvert > t_{\alpha/2,n-1}$	$2 \cdot tcdf(\lvert t_c \rvert, 999, n-1)$
			$H_1 : \mu_D > 0$		$t_c > t_{\alpha,n-1}$	$tcdf(\lvert t_c \rvert, 999, n-1)$
			$H_1 : \mu_D < 0$		$t_c < -t_{\alpha,n-1}$	
One-Sample Proportion		$H_0 : p = p_0$	$H_1 : p \neq p_0$	$z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\lvert z_c \rvert > z_{\alpha/2}$	$2 \cdot normalcdf(\lvert z_c \rvert, 999)$
			$H_1 : p > p_0$		$z_c > z_\alpha$	$normalcdf(\lvert z_c \rvert, 999)$
			$H_1 : p < p_0$		$z_c < -z_\alpha$	
Difference Between Two Proportions (Independent Samples)		$H_0 : p_1 = p_2$	$H_1 : p_1 \neq p_2$	$z_c = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$	$\lvert z_c \rvert > z_{\alpha/2}$	$2 \cdot normalcdf(\lvert z_c \rvert, 999)$
			$H_1 : p_1 > p_2$		$z_c > z_\alpha$	$normalcdf(\lvert z_c \rvert, 999)$
			$H_1 : p_1 < p_2$		$z_c < -z_\alpha$	