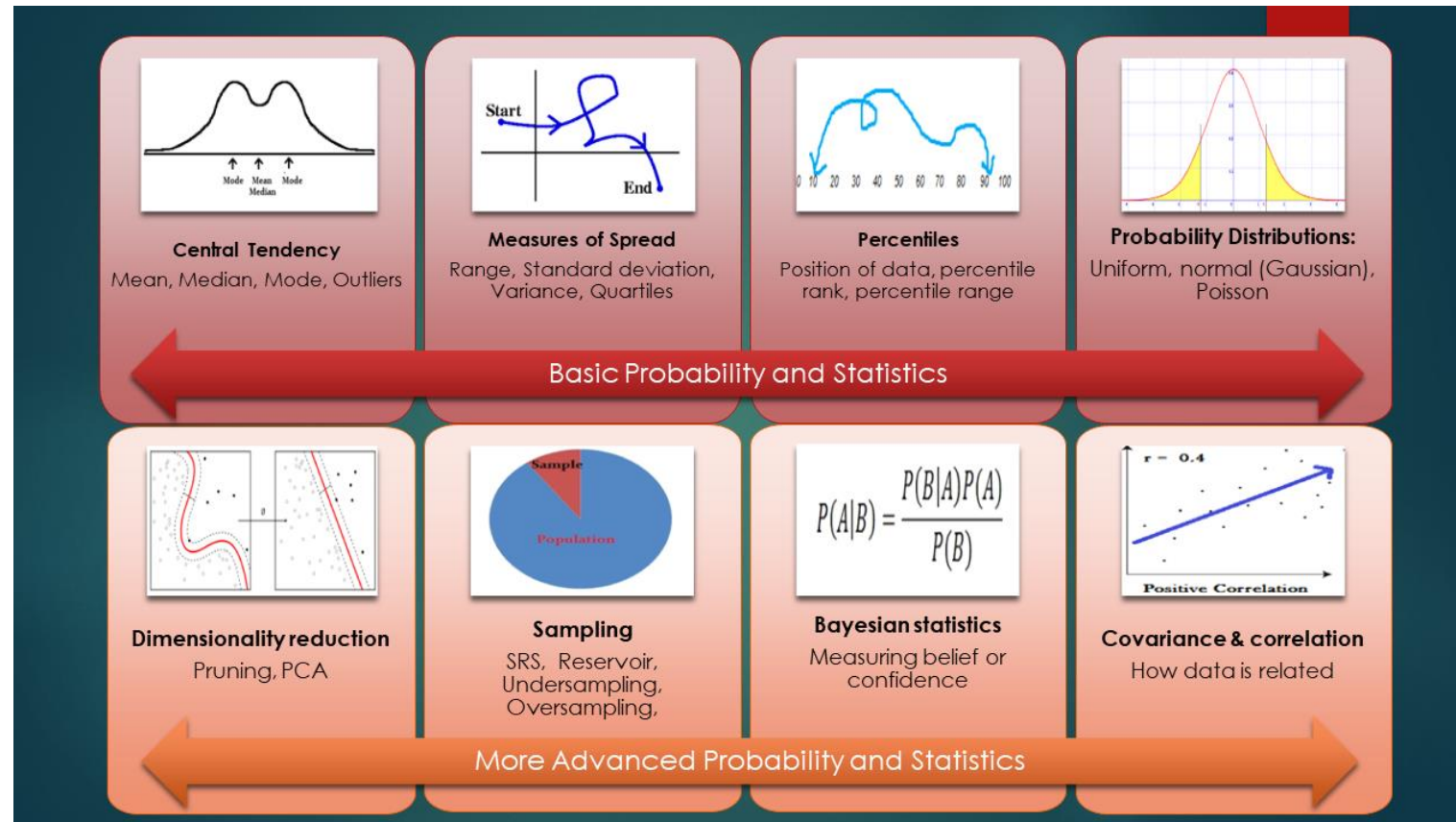


# Statistics Concepts for Data Scientists



# Population and Sample

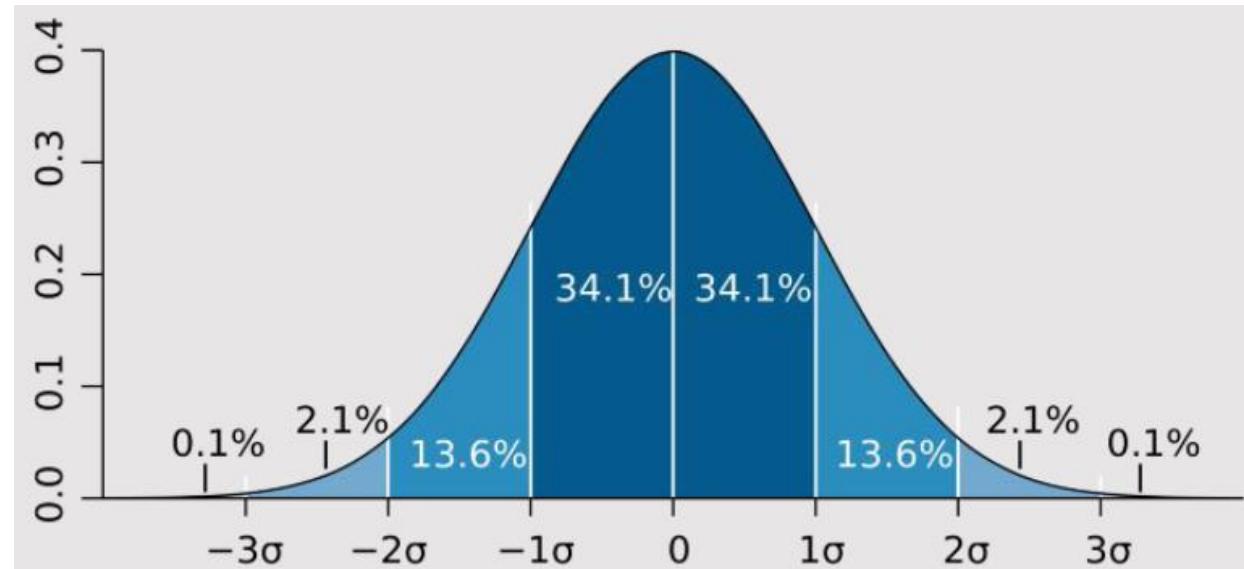
- Population is all elements in a group.
  - 25-year-old people in Europe is a population that includes all of the people that fits the description.
- It is not always feasible or possible to do an analysis on population because we cannot collect all the data of a population.
- Therefore, we use samples, which is a subset of a population.
  - 1000 college students in the US is a subset of the “college students in the US” population.
  - Sample’s characteristics are taken to be representative of the population.

# Normal Distribution

- A probability distribution is a function that shows the probabilities of the outcomes of an event or experiment.
- Probability distribution function of a variable shows the likelihood of the values it can take.
- Probability distribution functions are quite useful in predictive analytics or machine learning.
- We can make predictions about a population based on the probability distribution function of a sample from that population.

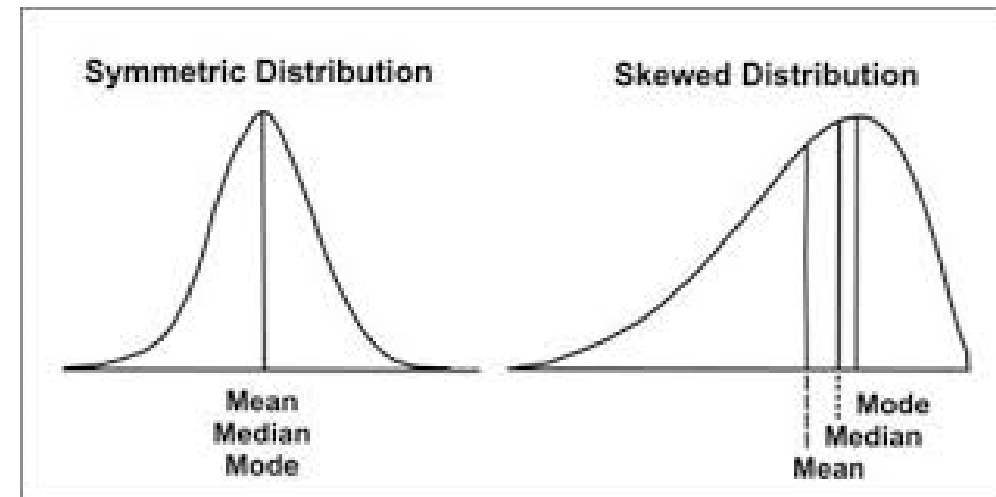
# Normal Distribution

- Normal distribution is a probability distribution function that looks like a bell.
- The peak of the curve indicates the most likely value the variable can take.
- The percentages indicate the % of data that falls in that region.
- As we move away from the mean, we start to see more extreme values with less probability observed.



# Measures of Central Tendency

- Central tendency is the central (or typical) value of a probability distribution.
- The most common measures of central tendency are mean, median, and mode.
- Mean is the average of the values in series ( $\mu$  for population and  $\bar{x}$  for the sample).
- Median is the value in the middle when values are sorted in order.
- Mode is the value that appears most often.



# Variance and Standard Deviation

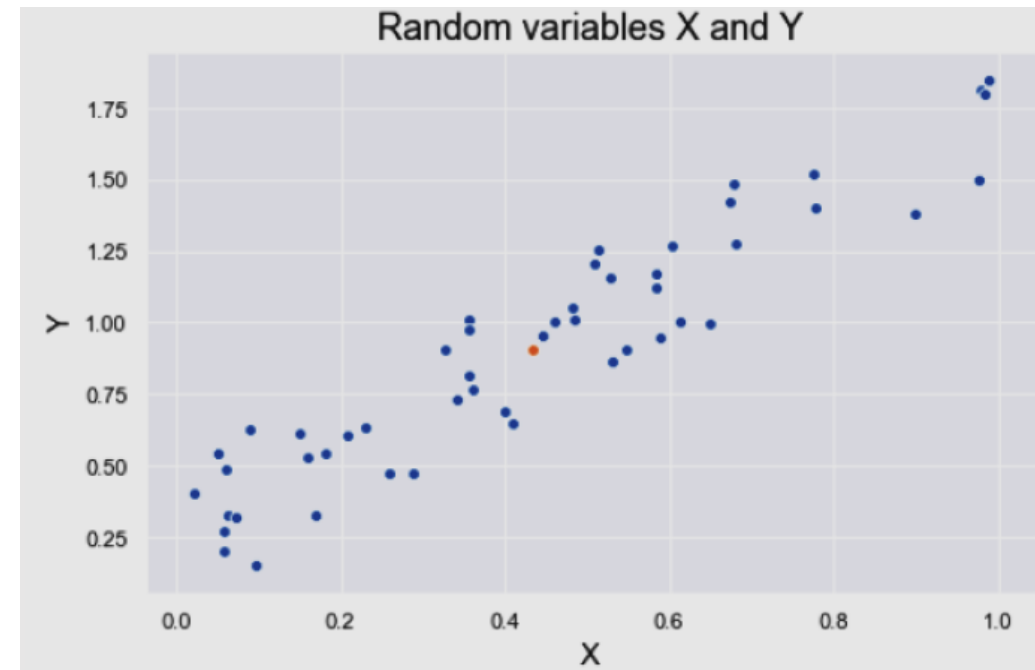
- Variance is a measure of the variation among values.
- It is calculated by adding up squared differences of each value and the mean and then dividing the sum by the number of samples.
- Standard deviation (STD) is a measure of how spread out the values are.
- STD is the square root of the variance.

$$\text{Variance} = \frac{\sum (x_i - \text{mean})^2}{N}$$

$$\text{Standard Deviation} = \sqrt{\frac{\sum (x_i - \text{mean})^2}{N}}$$

# Covariance and Correlation

- Covariance is a quantitative measure that represents how much the variations of two variables match each other.
- Covariance compares two variables in terms of the deviations from their mean (or expected) value.
- The orange dot represents the mean of these variables.
- The values change similarly with respect to the mean value of the variables.
- Thus, there is positive covariance between X and Y.



# Covariance and Correlation

- The formula for the covariance of two random variables:

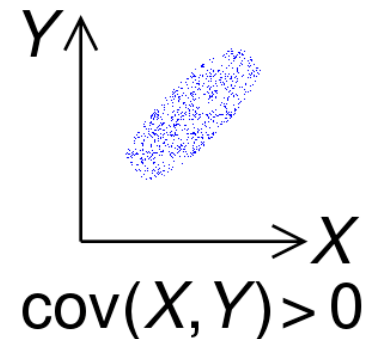
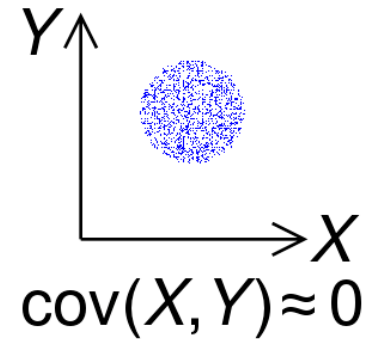
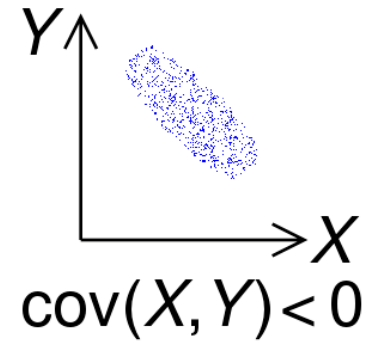
For Population

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

Note: The covariance of a variable with itself is the variance of that variable.





# Covariance and Correlation

- Correlation is a normalization of covariance by the standard deviation of each variable.
- Covariance indicates the direction of the linear relationship between variables while correlation measures both the strength and direction of the linear relationship between two variables.
- When you divide the covariance values by the standard deviation, it essentially scales the value down to a limited range of -1 to +1.

$$\text{Correlation} = \frac{\text{Cov}(x, y)}{\sigma x * \sigma y}$$

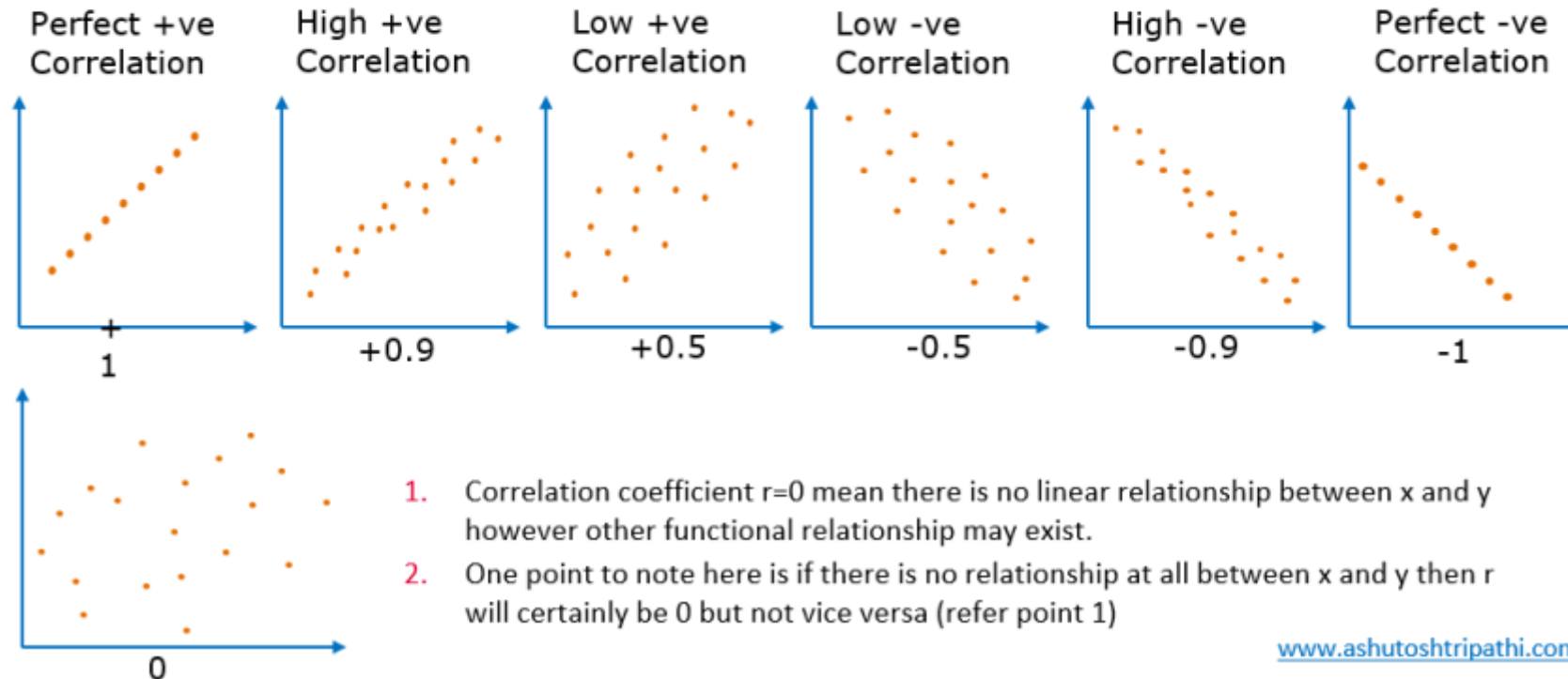
# Covariance and Correlation

Correlation coefficient  $r$  is number between -1 to +1 and tells us how well a regression line fits the data and defined by

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where,

- $s_{xy}$  is the covariance between  $x$  and  $y$
- $s_x$  and  $s_y$  are the standard deviations of  $x$  and  $y$  respectively.

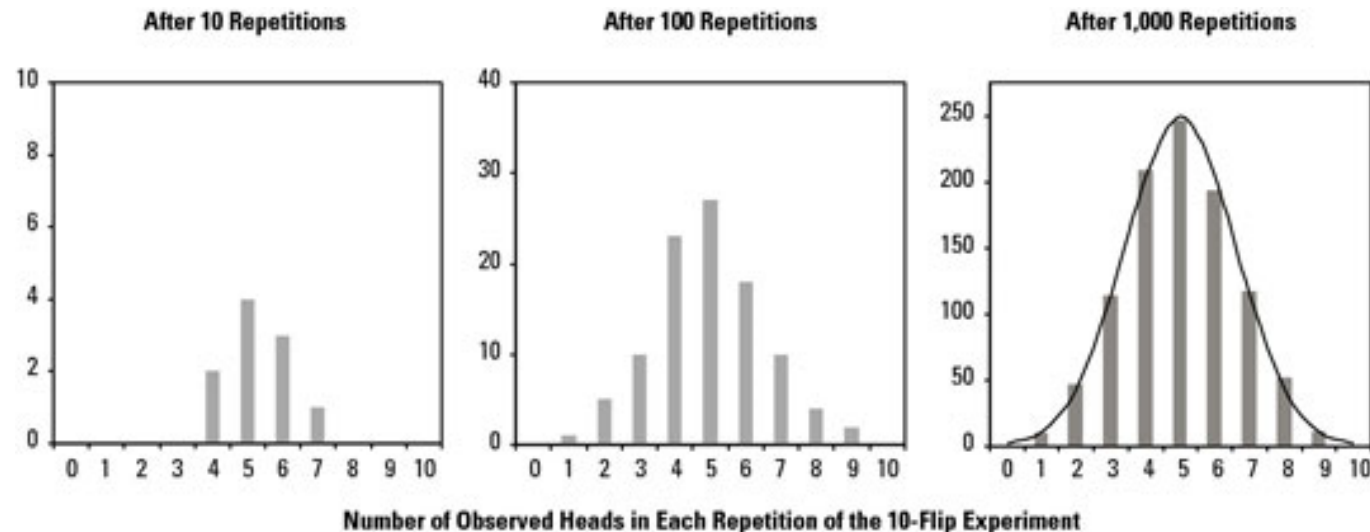


# Covariance & Correlation

CORRELATION	COVARIANCE
There is said to be correlation between two, when change in one results in change in another.	Covariance talks about the direction of the relationship between the two variables (positive or negative)
CORRELATION	COVARIANCE
Measures the strength of the variables under comparison	Measures the extent of change in one with regards to change in another.
Correlation is a scaled down version of covariance.	Covariance is considered as a part of correlation.
Value here lies between -1 and +1.	Value here lies between -infinity to +infinity
Correlation is a unit-free measure	Covariance value is the product of the units of the variables.
There would be no change in correlation due to scale.	Any change in scale affects covariance.

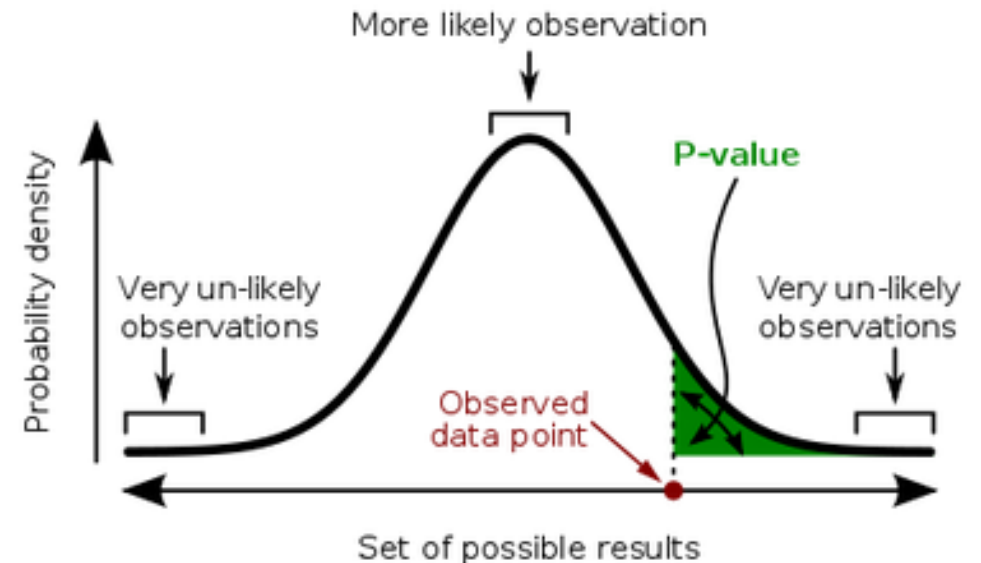
# Central Limit Theorem

- In many fields, including natural and social sciences, when the distribution of a random variable is unknown, a normal distribution is used.
- The central limit theorem (CLT) justifies why normal distribution can be used in such cases.
- According to CLT, as we take more samples from a distribution, the sample averages will tend towards a normal distribution regardless of the population distribution.



# P-value

- P-value is a measure of the likelihood of a value that a random variable takes.
- In null hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the results actually observed, assuming that the null hypothesis is correct.
- A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.
- A p-value of 0.05 means that there is a 5 % chance that the results are due to random chance.



# Hypothesis Testing - Example

Hypothesis Testing, a 5-step approach using the traditional method

Using a 0.05 significance level, we are testing the claim that more than half (of all Americans) admit to running red lights. We have a sample of 880 randomly selected drivers, of which 56% admit they run red lights. Obviously, in the study, 56% is more than 50%, but is it significantly more than half? A hypothesis test is a test of significance.

The following are five steps that will lead you to correct conclusions concerning the null hypothesis and the claim. Instructor comments as we go along are in parentheses.

- [1] Claim:  $p > .5$  (The claim will appear as the null or the alternative hypothesis.)  
 $H_0$ :  $p = .5$  (The null hypothesis always contains the condition of equality (=) .  
 $H_1$ :  $p > .5$  (The alternative hypothesis is the same as the claim.)

- [2] (Sketch and label critical value. Look at the direction of the inequality symbol in  $H_1$  to determine where to shade. ).



- [3] Decision Rule:  
We will reject  $H_0$  if  $z_{ts} > 1.645$ .

(This step explains to someone following the logic what the plan is. We will always reject the null hypothesis if the test statistic,  $z_{ts}$ , falls in the critical region (shaded).)

[4] 
$$z_{ts} = \frac{0.56 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{880}}} = 3.56$$

(Calculate the test statistic)

- [5] (State the conclusion in two parts: 1) about the null hypothesis and 2) about the claim. The flowchart on the Triola formula card will help you discuss the claim correctly. We said in the Decision Test (step 3) that we would reject the null hypothesis if the test statistic is greater than 1.645. It is. It is 3.56. Conclusions 1) and 2) follow from that observation.)

- 1) Reject  $H_0$ .  
2) The sample data supports the claim that most Americans admit to running red lights.

# Bayes' Theorem

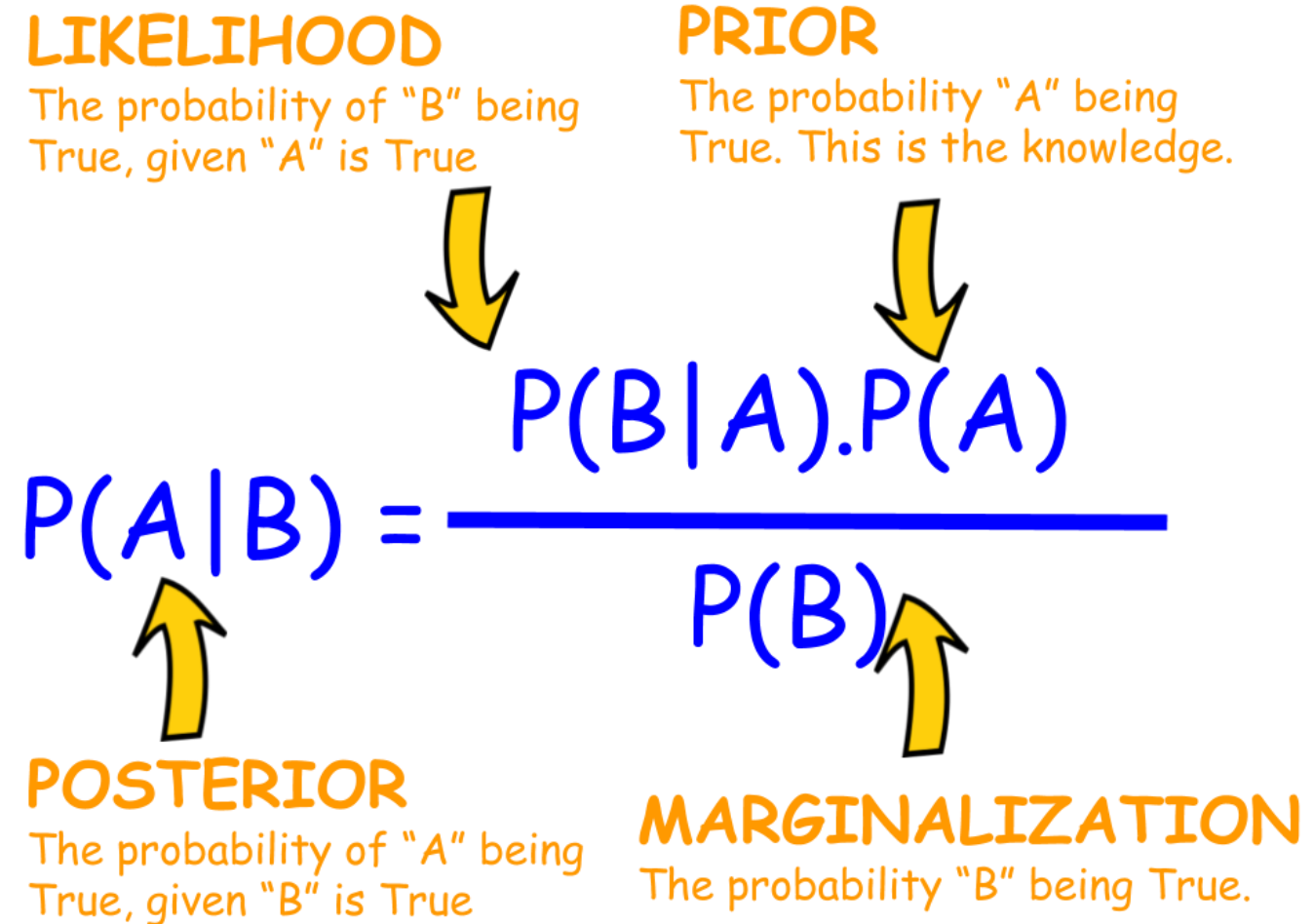
- According to Bayes' theorem, the probability of event A given that event B has already occurred can be calculated using the probabilities of event A and event B and probability of event B given that A has already occurred.
- Bayes' theorem is so fundamental and ubiquitous that a field called "Bayesian statistics" exists.

## LIKELIHOOD

The probability of "B" being True, given "A" is True

## PRIOR

The probability "A" being True. This is the knowledge.


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

The diagram shows the formula for Bayes' Theorem. Four yellow arrows point from the component labels to the formula: one from 'LIKELIHOOD' to  $P(B|A)$ , one from 'PRIOR' to  $P(A)$ , one from 'POSTERIOR' to  $P(A|B)$ , and one from 'MARGINALIZATION' to  $P(B)$ .

## POSTERIOR

The probability of "A" being True, given "B" is True

## MARGINALIZATION

The probability "B" being True.