1. What's Statistical Modeling?

You develop mathematical equations that describe the interrelation between two or more variables.

2. Statistical Analysis v/s Non Statistical analysis

Statistical (Quantitative): Exploring large data to identify patterns, trends, relationship between variables.
NonStatistical (Qualitative): Generic info. Includes text, images, audio etc

3. Descriptive v/s Inferential Statistics

Descriptive: Summarizes the data numerically or graphically. average, mean, mode. std dev, correlation
Inferential : Generalizes large datasets and applies probability theory to draw a conclusion. Used to model relationships between the variables in the data. Eg. Hypothesis testing

4. Types of Descriptive Statistical measures (4 types)

a. Measure of frequency (Numbers, percentage)
b. Measure of central tendency (Mean, Median, Mode)
c. Measure of Spread/ Dispersion (Std deviation, Variance)
d. Measure of Positions (Percentiles, quartiles)

5. Variance v/s Std Deviation

The standard deviation is a measure of spread or dispersion of data around its center. A deviation is the distance from an observation to its mean. The bigger the deviations, the more spread there is. However, deviations can be either positive (greater than the mean) or negative (less than the mean). So the deviation is difficult to use to measure overall dispersion because the positives and negatives cancel each other out (the sum of all the deviations from the mean is zero). So we square the deviation to remove the sign and then add them up. This is the sum of squared deviations. Taking the average (divide by the number of observations) gives the mean of the squared deviations, also called the variance. We take the average because the sum of squared deviations will be a function of sample

size, more deviations makes the sum of squared deviations bigger. But the variance is in the square of the original units.
The standard deviation is the square root of the variance.

6. Types of variables

a. Nominal: They have two or more categories. Eg. blood group, Gender
b. Ordinal: They have values in logical order. Relative distance between two data values is not clear. Eg. size of cup- small, medium, large
c. Interval: They have equal difference between scale values that have quantitative meaning. The interval scale does not have a true zero point. Eg. Fahrenheit degree scale for temperature
d. Ratio: Same as interval but they have zero point. Eg. System of inches used with a common ruler.

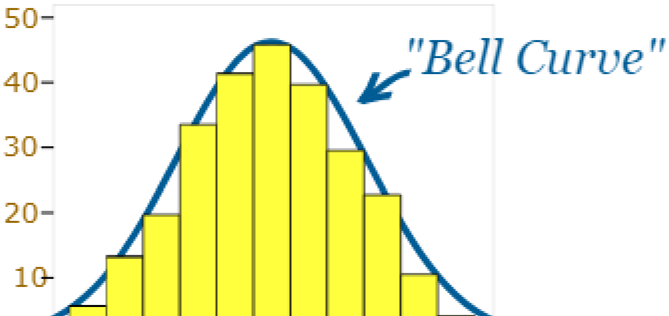
7. Sampling and its types (4 types)

a. Random
b. Systematic - Every member of pop. is numbered and every Kth member is selected.
c. Stratified - pop. is divided into groups/ strata based on characteristics (gender, age group, income). Then you randomly select from every strata.
d. Cluster - pop. is divided into subgroups. Each subgroup should have the same qualities. Then you randomly select the entire subgroups.

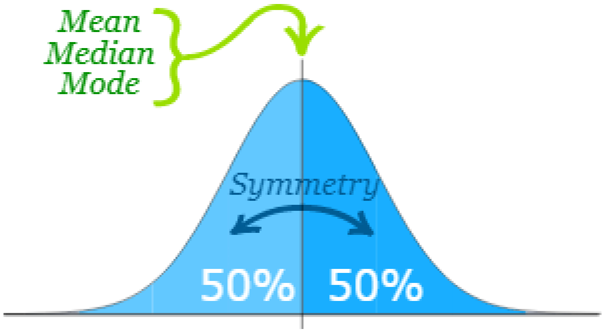
8. Normal Distribution

The normal distribution is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

Data can be "distributed" (spread out) in different ways.
It can be spread out more on the left or more on the right or it can be all jumbled up data random.
But there are many cases where the data tends to be around a central value with no bias left or right, and it gets close to a "Normal Distribution" like this:



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

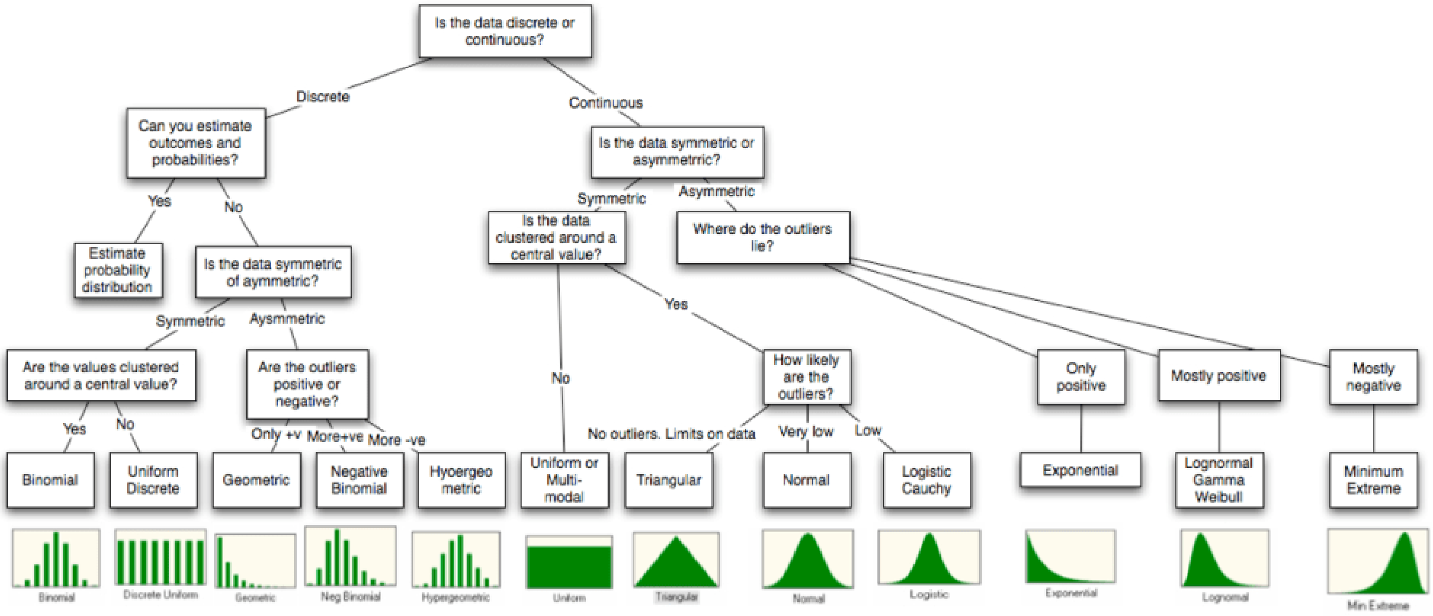


The Normal Distribution has:

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

9. All graphs

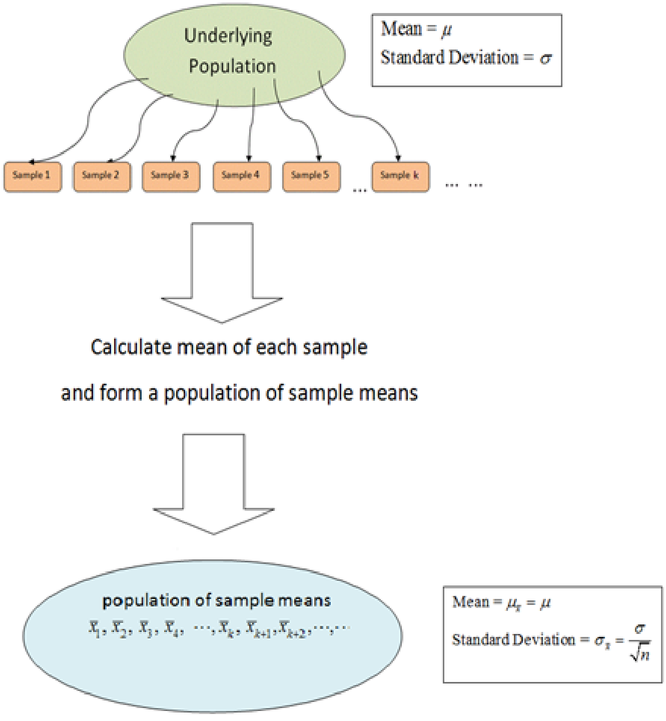
Figure 6A.15: Distributional Choices

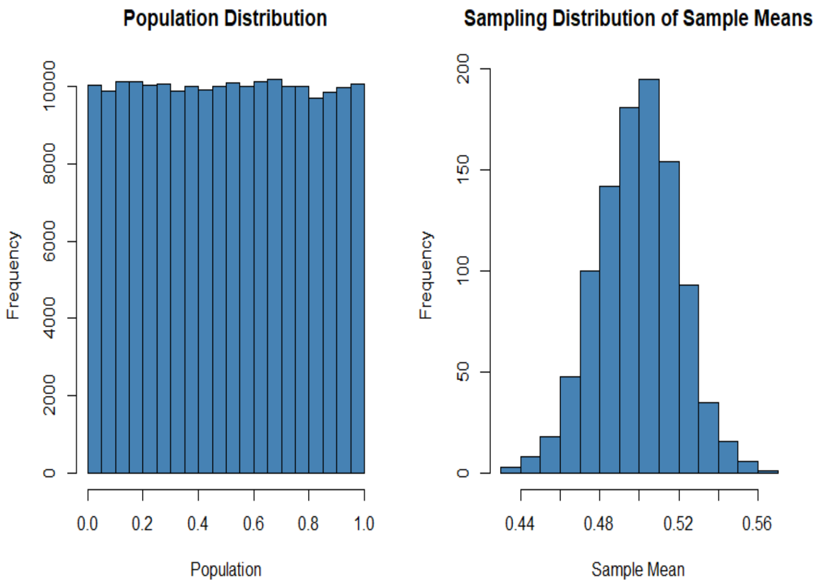


10. Central Limit theorem

For a population with any distribution, the Sampling distribution of sample means approaches Normal distribution as the sample size increases.

When you take the mean of ‘n’ no. of sample and plot the, , they will represent a bell curve (Normal distribution). As you increase sample size, the bell curve becomes thin.


Underlying Population
Mean = μ
Standard Deviation = σ
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
...
Sample k
... ...
Calculate mean of each sample
and form a population of sample means
population of sample means
$\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \cdots, \bar{x}_k, \bar{x}_{k+1}, \bar{x}_{k+2}, \cdots, \cdots$
Mean = $\mu_{\bar{x}} = \mu$
Standard Deviation = $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$



11. Hypothesis Testing (H0, Ha, P-value, alpha, CI, degrees of freedom, which test, errors)

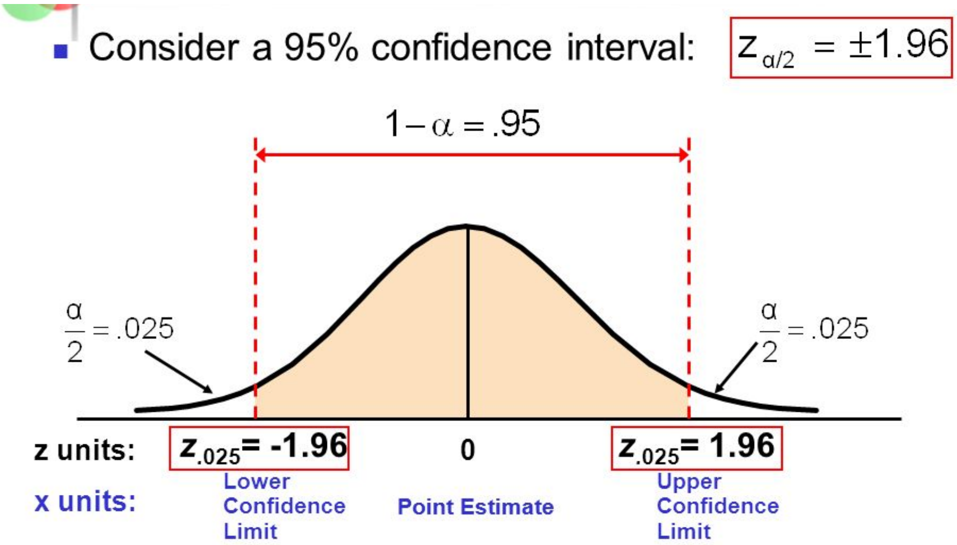
Terminology

- **Null Hypothesis:** the hypothesis that sample observations result purely from chance. The null hypothesis tends to state that there's no change.
- **Alternative Hypothesis:** the hypothesis that sample observations are influenced by some non-random cause.
- **P-value:** the probability of obtaining the observed results of a test, assuming that the null hypothesis is correct; a smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.
- **Alpha:** the significance level; the probability of rejecting the null hypothesis when it is true — also known as **Type 1 error.**
- **Degrees of freedom:** It's an estimate is the number of independent pieces of information that went into calculating the estimate. It's not quite the same as the number of items in

the sample. In order to get the df for the estimate, you have to subtract 1 from the number of items.

- **Confidence interval & levels:**

  Confidence levels are expressed as a percentage (for example, a 95% confidence level). It means that should you repeat an experiment or survey over and over again, 95 percent of the time your results will match the results you get from a population (in other words, your statistics would be sound!). Confidence intervals are your results and they are usually numbers. For example, you survey a group of pet owners to see how many cans of dog food they purchase a year. You test your statistic at the 99 percent confidence level and get a confidence interval of (200,300). That means you think they buy between 200 and 300 cans a year. You're super confident (99% is a very high level!) that your results are sound, statistically.



Reject or Do not Reject?

If the **P**-value is **G**reater than the **A**lpha, **D**o not **R**eject the **N**ull.

What is the point of Significance Testing?

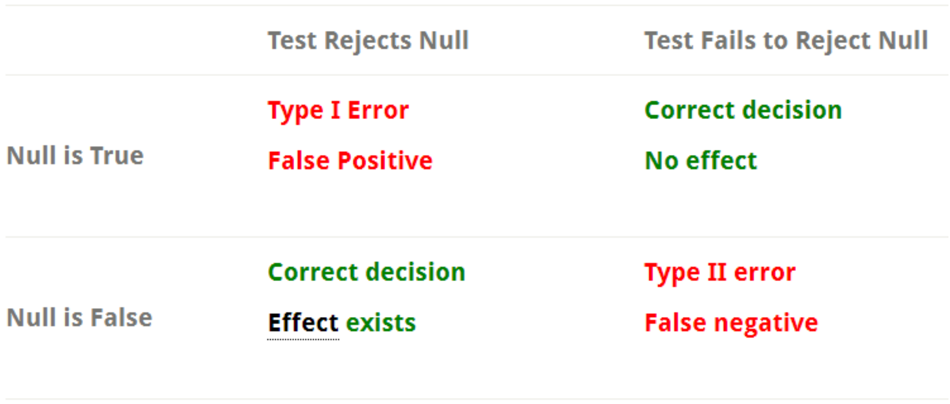
It's used to determine how likely or unlikely a hypothesis is for a given sample of data. The last part of the statement, 'for a given sample of data' is key because more often than not, you won't be able to get an infinite amount of data or data that represents the entire population.

Steps for Hypothesis Testing

Here are the steps to performing a hypothesis test:

1. State your null and alternative hypotheses. To reiterate, the null hypothesis typically states that everything is as normally was — that nothing has changed.
2. Set your significance level, the alpha. This is typically set at 5% but can be set at other levels depending on the situation and how severe it is to commit a type 1 and/or 2 error.
3. Collect sample data and calculate sample statistics.
4. Calculate the p-value given sample statistics. Once you get the sample statistics, you can determine the p-value through different methods. The most common methods are the T-score and Z-score for normal distributions.
5. Reject or do not reject the null hypothesis.

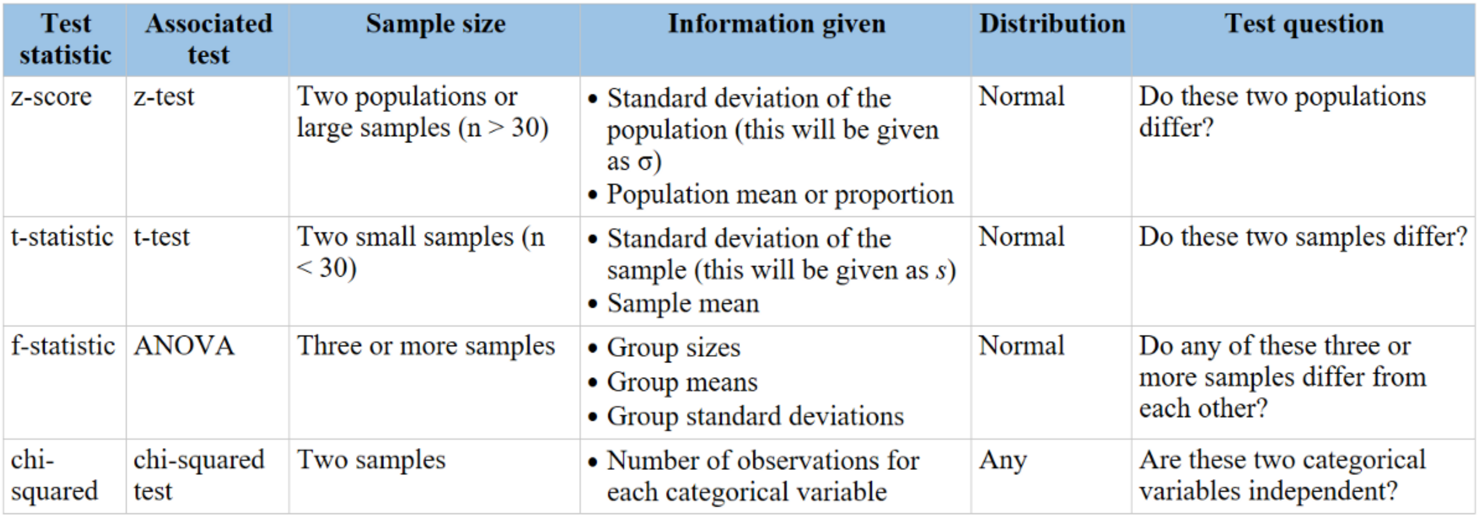
Errors



	Test Rejects Null	Test Fails to Reject Null
Null is True	Type I Error False Positive	Correct decision No effect
Null is False	Correct decision Effect exists	Type II error False negative

Type 1: the test result says Netflix new feature is working, but actually it isn't.
Type 2: the test result says you don’t have coronavirus, but you actually do.

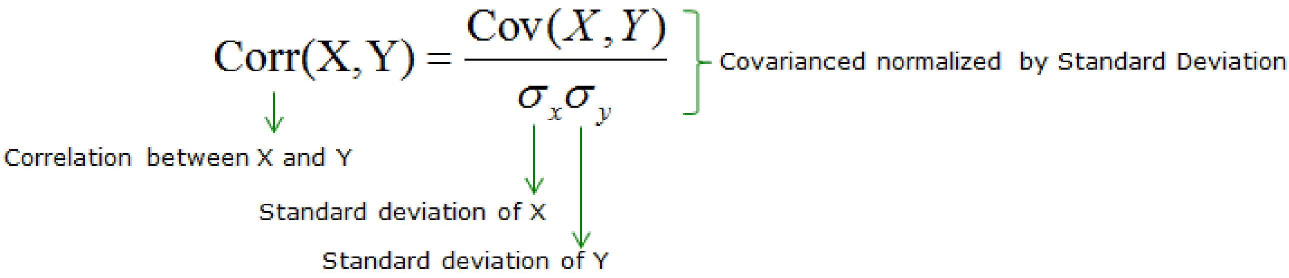
12. Tests in Hypothesis testing (chi sq, t-test, Anova)



Test statistic	Associated test	Sample size	Information given	Distribution	Test question
z-score	z-test	Two populations or large samples (n > 30)	• Standard deviation of the population (this will be given as σ) • Population mean or proportion	Normal	Do these two populations differ?
t-statistic	t-test	Two small samples (n < 30)	• Standard deviation of the sample (this will be given as *s*) • Sample mean	Normal	Do these two samples differ?
f-statistic	ANOVA	Three or more samples	• Group sizes • Group means • Group standard deviations	Normal	Do any of these three or more samples differ from each other?
chi-squared	chi-squared test	Two samples	• Number of observations for each categorical variable	Any	Are these two categorical variables independent?

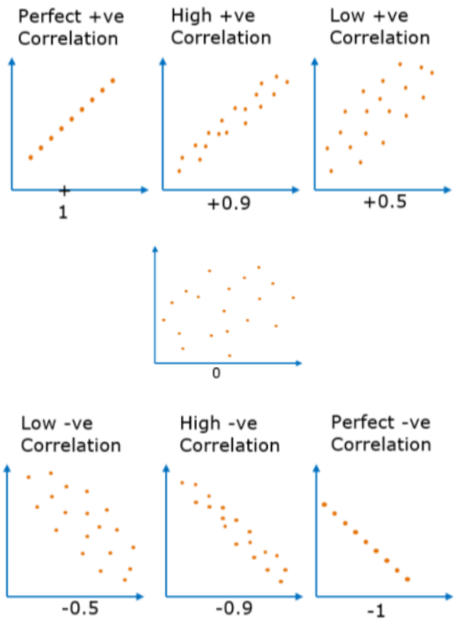
13. Covariance and Correlation

Covariance is nothing but a measure of correlation. Correlation refers to the scaled form of covariance. Covariance indicates the direction of the linear relationship between variables. Correlation on the other hand measures both the strength (how much they are dependent on one another) and direction of the linear relationship between two variables.



$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Covarianced normalized by Standard Deviation
Correlation between X and Y
Standard deviation of X
Standard deviation of Y



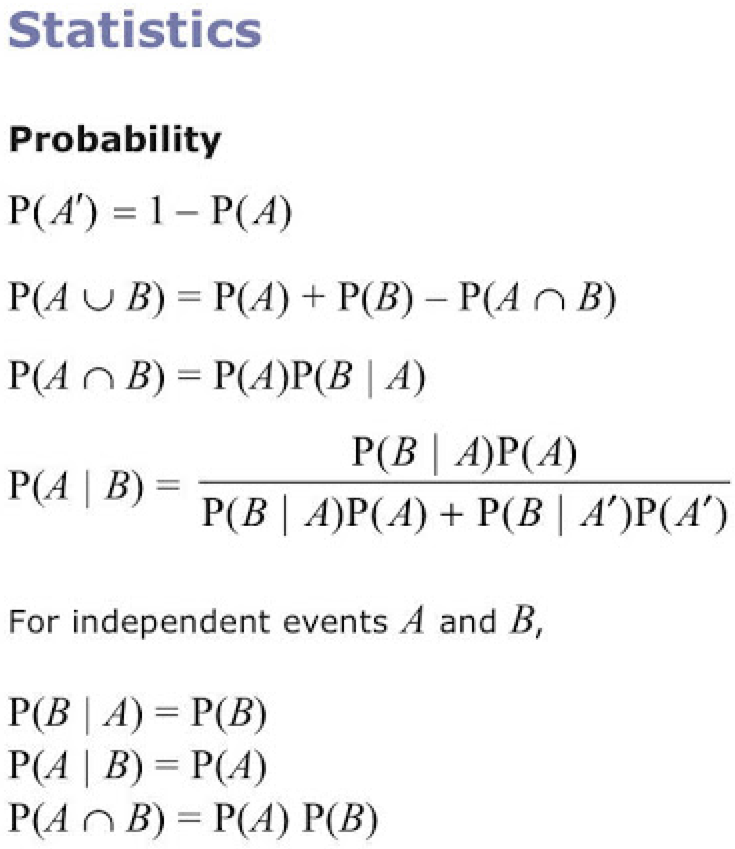
14. Random Variable

A **Random Variable** is a set of possible values from a random experiment.

15. Expected values of Random variable

The **expected value** is simply a way to describe the average of a discrete set of variables based on their associated probabilities. This is also known as a probability-weighted average. For this example, it would be estimated that you would work out 2.1 times in a week, 21 times in 10 weeks, 210 times in 100 weeks, etc.
In statistics and probability analysis, the expected value is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all of those values.

16. Basic Probability



Statistics

Probability

$$P(A') = 1 - P(A)$$

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

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

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

For independent events A and B,

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

17. Conditional Probability

Conditional probability refers to the chances that some outcome occurs given that another event has also occurred.

It is often stated as the probability of B given A and is written as P(B|A), where the probability of B depends on that of A happening.

The formula for conditional probability is:

P(B|A) = P(A and B) / P(A)

which you can also rewrite as:

P(B|A) = P(A∩B) / P(A)

Finding Hidden Data

Using Algebra we can also "change the subject" of the formula, like this:

Start with:	P(A and B) = P(A) x P(B|A)
Swap sides:	P(A) x P(B|A) = P(A and B)
Divide by P(A):	P(B|A) = P(A and B) / P(A)

And we have another useful formula:

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

*"The probability of **event B given event A** equals*
*the probability of **event A and event B** divided by the probability of **event A***

Example: Ice Cream

70% of your friends like Chocolate, and 35% like Chocolate AND like Strawberry.

What percent of those who like Chocolate also like Strawberry?

P(Strawberry|Chocolate) = P(Chocolate and Strawberry) / P(Chocolate)

→ 0.35 / 0.7 = 50%

50% of your friends who like Chocolate also like Strawberry

Big Example: Soccer Game

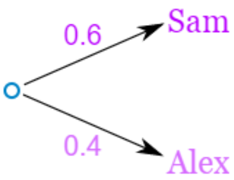
You are off to soccer, and want to be the Goalkeeper, but that depends who is the Coach today:

- with Coach Sam the probability of being Goalkeeper is **0.5**
- with Coach Alex the probability of being Goalkeeper is **0.3**

Sam is Coach more often ... about 6 out of every 10 games (a probability of **0.6**).

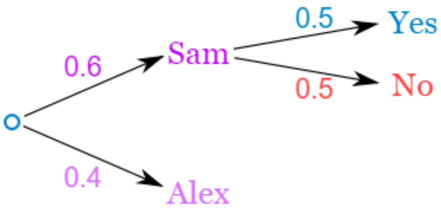
So, what is the probability you will be a Goalkeeper today?

Let's build a tree diagram. First we show the two possible coaches: Sam or Alex:

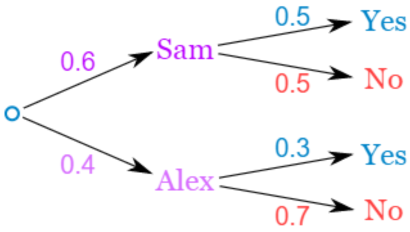


The probability of getting Sam is 0.6, so the probability of Alex must be 0.4 (together the probability is 1)

Now, if you get Sam, there is 0.5 probability of being Goalie (and 0.5 of not being Goalie):



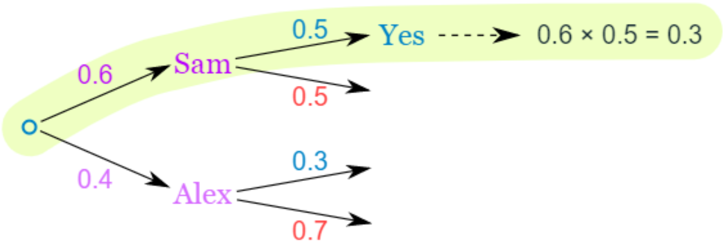
If you get Alex, there is 0.3 probability of being Goalie (and 0.7 not):



The tree diagram is complete, now let's calculate the overall probabilities. Remember that:

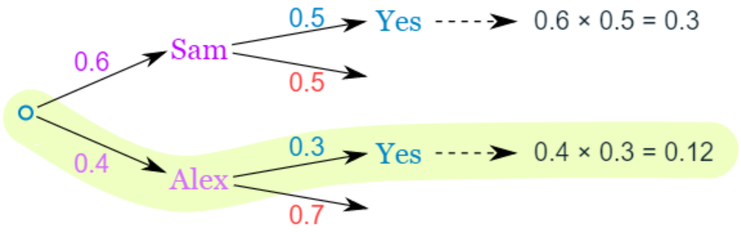
P(A and B) = P(A) x P(B|A)

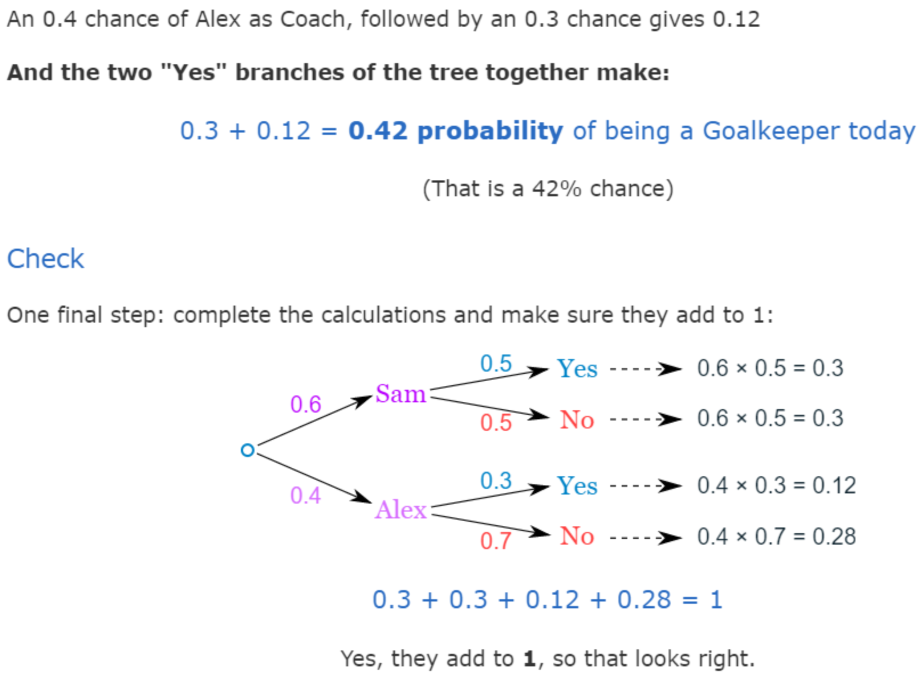
Here is how to do it for the "Sam, Yes" branch:



(When we take the 0.6 chance of Sam being coach times the 0.5 chance that Sam will let you be Goalkeeper we end up with an 0.3 chance.)

But we are not done yet! We haven't included Alex as Coach:





An 0.4 chance of Alex as Coach, followed by an 0.3 chance gives 0.12

And the two "Yes" branches of the tree together make:

0.3 + 0.12 = **0.42 probability** of being a Goalkeeper today

(That is a 42% chance)

Check

One final step: complete the calculations and make sure they add to 1:

0.3 + 0.3 + 0.12 + 0.28 = 1

Yes, they add to **1**, so that looks right.

18. Bayes’ theorem

Ever wondered how computers learn about people?

Example:

An internet search for "movie automatic shoe laces" brings up "Back to the future"

Has the search engine watched the movie? No, but it knows from lots of other searches what people are **probably** looking for.

And it calculates that probability using Bayes' Theorem.

Bayes' Theorem is a way of finding a probability when we know certain other probabilities.

The formula is:

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

Which tells us: how often A happens *given that B happens*, written **P(A|B)**,

When we know: how often B happens *given that A happens*, written **P(B|A)**

and how likely A is on its own, written **P(A)**

and how likely B is on its own, written **P(B)**

Let us say P(Fire) means how often there is fire, and P(Smoke) means how often we see smoke, then:

P(Fire|Smoke) means how often there is fire when we can see smoke
P(Smoke|Fire) means how often we can see smoke when there is fire

So the formula kind of tells us "forwards" P(Fire|Smoke) when we know "backwards" P(Smoke|Fire)

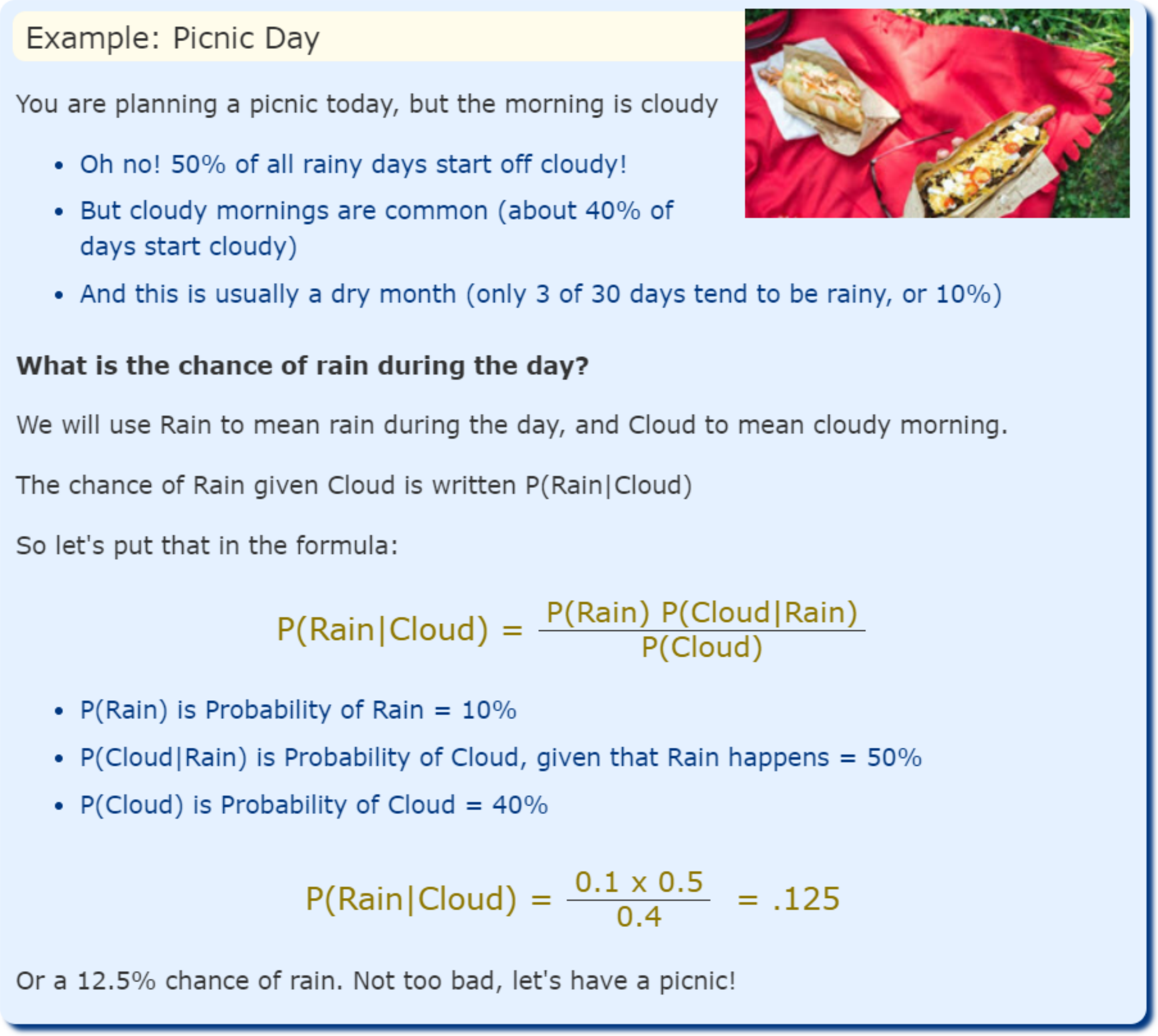
Example:

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$\begin{aligned} \text{P(Fire|Smoke)} &= \frac{\text{P(Fire) P(Smoke|Fire)}}{\text{P(Smoke)}} \\ &= \frac{1\% \text{ x } 90\%}{10\%} \\ &= 9\% \end{aligned}$$

So it is still worth checking out any smoke to be sure.



Example: Picnic Day

You are planning a picnic today, but the morning is cloudy

- Oh no! 50% of all rainy days start off cloudy!
- But cloudy mornings are common (about 40% of days start cloudy)
- And this is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

We will use Rain to mean rain during the day, and Cloud to mean cloudy morning.

The chance of Rain given Cloud is written P(Rain|Cloud)

So let's put that in the formula:

$$\text{P(Rain|Cloud)} = \frac{\text{P(Rain) P(Cloud|Rain)}}{\text{P(Cloud)}}$$

- P(Rain) is Probability of Rain = 10%
- P(Cloud|Rain) is Probability of Cloud, given that Rain happens = 50%
- P(Cloud) is Probability of Cloud = 40%

$$\text{P(Rain|Cloud)} = \frac{0.1 \text{ x } 0.5}{0.4} = .125$$

Or a 12.5% chance of rain. Not too bad, let's have a picnic!