Probabilities Cheat Sheet

Probabilities... (A.K.A. Chance)

How likely something (an event) is to happen

Kinds of Probabilities

Conditional Probabilities – Probability of an event happening based on whether or not something else happened

Join Probabilities – Probability of two events happening at the same time

Marginal Probabilities – A.K.A. Unconditional Probabilities, are just the summation of all probabilities

$$\textit{Probability} = \frac{\textit{How many times event happened}}{\textit{Total Outcomes}}$$

Kinds of Events

Mutually Exclusive – Events that can't happen at same time

Non-Mutually Exclusive – Events that can happen at same time

Independent – When an event's probability isn't affected by anything else happening or not happening (e.g. a coin toss isn't affected by previous coin toss)

Dependent – Events whose probabilities change based on each other happening or not happening

Remember! Probabilities are always between 0 and 1, if you get a probability outside this range there is something wrong with calculation.

0 = Not gonna happen

1 = Definitely gonna happen

Big Takeaway

Probabilities can give you an indication of what is likely to happen, but t**hey CANNOT tell you what *will* happen.**

Here's an example… probability tells us that when you flip a coin you have a 50% chance of hitting either heads or tails, so if you flip the coin 100 times you would expect to get 50 heads and 50 tails… but that isn't always what will happen!

All that probability is telling us is that most of the time you'll get a number somewhere close to 50 heads and 50 tails, not that you absolutely will get that exact number each and every time you run the "100 coin flip test."

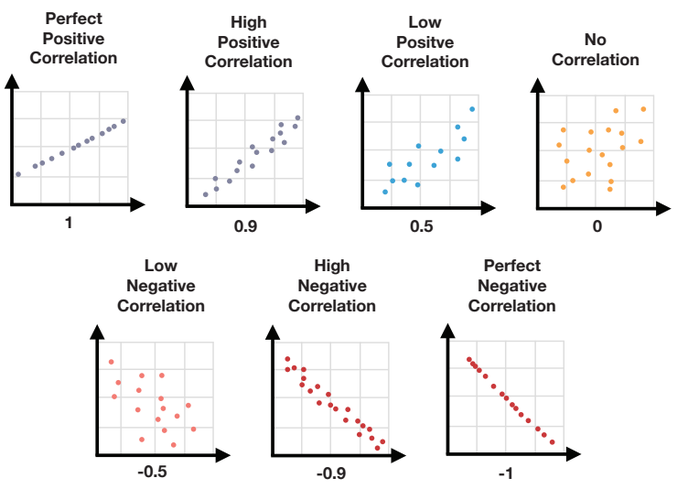
Correlation

When there is some relationship between two things

- Correlations always take values between -1 and 1

> **-1 is a perfect negative correlation**, which means as one thing gets bigger the other thing gets smaller
>
> **0 is no correlation at all**, basically there is no relationship between these things
>
> **1 is a perfect positive correlation**, which means that when one thing gets bigger so does the other

- The closer the correlation value is to -1 or 1, the tighter (more linear) the relationship will be on a scatter plot (see below on Pearson's coefficient)



Some different correlation calculations...

There are different correlation calculations (called coefficients) for different kinds of data:

- **Pearson's Coefficient** – Measures linear relationship between two variables
- **Spearman's Rank Coefficient** – Measures relationship between two ordinal variables
- **Phi Coefficient** – Measures relationship between two dichotomous variables

Pearson's Coefficient is most popular and what analytical tools use.

Here's how we calculate correlation (Pearson's way):

In this example we have two things to compare, **X** and **Y.**

1. First calculate Mean (average) of **X**
2. Calculate Mean (average) of **Y**
3. Subtract Mean of **X** from each of **X** values (we'll call these **A**), and subtract Mean of **Y** from each of **Y** values (we'll call these **B**)
4. Square **A's** (we'll call these **C^2's**)
5. Square **B's** (we'll call these **D^2's**)
6. Multiply all **A's** by **B's** (we'll call these **AB's**)
7. Add up all **AB's**
8. Add up all **C^2's**
9. Add up all **D^2's**
10. Now perform calculation below…

$$\textbf{Correlation} = \frac{\textit{Sum of all AB's}}{\sqrt{\textit{(Sum of C2's) x (Sum of D2's)}}}$$

Definitions

Coefficient – Basically just a static numerical value that is used in a calculation

Linear – Like, or in shape of, a line

Variable – Something we take into account in our analysis

Ordinal Data – Data that is ordered so that its values indicate rank

Dichotomous Data – Data that takes on two values only (e.g. 1 or 0, True or False, Yes or No)

Logical Fallacies to Avoid...

...to make better arguments

Fallacy – An error in reasoning that will undermine your argument

Slippery Slope – Saying that if A happens, and B-Y happen, then Z will happen, so basically A = Z

Hasty Generalizations – Just what it sounds like, you come to a conclusion about something before you have sufficient information

Post hoc ergo propter hoc – Basically saying that if B happens after A, then A caused B

Genetic – Saying that the origin of a person or a thing dictates its character or worth

Begging the Question - When you try to validate your conclusion within the question that you are asking

Circular Arguments – Stating a conclusion that just restates itself as proof (ex. My car is awesome because it's so cool)

Either-Or – Oversimplifying a conclusion by assuming that it must be either one thing or the other

⚠ Caution Hazard ⚠

Correlation DOES NOT prove Causation!

Beware temptation to say that a *correlation* between two things means one causes the other.

For example...
There may be a correlation between sweater and snow-shovel sales. However, that does not mean that sweaters make people buy snow-shovels.

All that we can say with a correlation is that there is a relationship/link between sweater sales and snow-shovel sales.

Multiple & Conditional Probability

Make sure that you account for ALL possible events when calculating probability.

Same is true for conditional vs. unconditional probabilities, be sure you understand all relationships.

Real life is never as simple as a coin toss.

Probabilities ARE NOT Guarantees!

Probability tells you that **over the long run** there is a certain chance of something happening. Not that something will or will not happen at a specific time.

In other words, **probabilities are great for general predictions about long term events, but they cannot and do not predict specific events.**

Over Generalization of Results

If you calculate a correlation on a specific population, you cannot then say that correlation is same for general population.

Make sure to review ⚠ Hazards! section regarding correlation and causation!

Check out our Statistics Cheat Sheet