Introduction to Probability

Probability - Why is it important?

The data generation process is highly uncertain, but understanding probability can help to deal with this uncertainty.

Data is always imperfect, and to draw inferences from this imperfect data, we treat the imperfections as the result of a random process

Probability also helps us to understand how good our inferences are - estimates will always have a probability attached or a margin of error, which can be determined by understanding probability.

Probability

Most processes are **stochastic** rather than **deterministic**. That is, the same inputs do not always produce the same outputs.

Why? The process could just be inherently random, or there are additional factors that we have not measured or have not considered that contribute to the different outcomes.

Probability - Informal Example

We can reason about random processes by using probability.

For example, we can reason about the possible outcome of a die roll using probability.

Probability - Informal Example

What is the chance of rolling a 5?

Answer: There are 6 equally-likely faces, and only one of them is 5, so the chances of rolling a 5 are 1/6.

Probability - Informal Example

What is the chance of rolling an even number?

Answer: There are 6 equally-likely faces, and three of them are even (2, 4, and 6), so the chances are 3/6.

Probability - Informal Example

What is the chance of rolling the die twice and it landing on a 4 both times?

Answer: We know that 1/6 of the time, it will land on 4 on the first roll and 1/6 of the time, it will and on 4 on the second roll.
Combining these, we can see that it will land on 4 both times (⅙)*(⅙) = 1/36 of the time.

Probability

Now, let’s get a little more formal in talking about probability.

If you want to get very formal about it, there is a whole branch of mathematics dedicated to studying probability, starting from the framework of Kolmogorov’s Axioms.

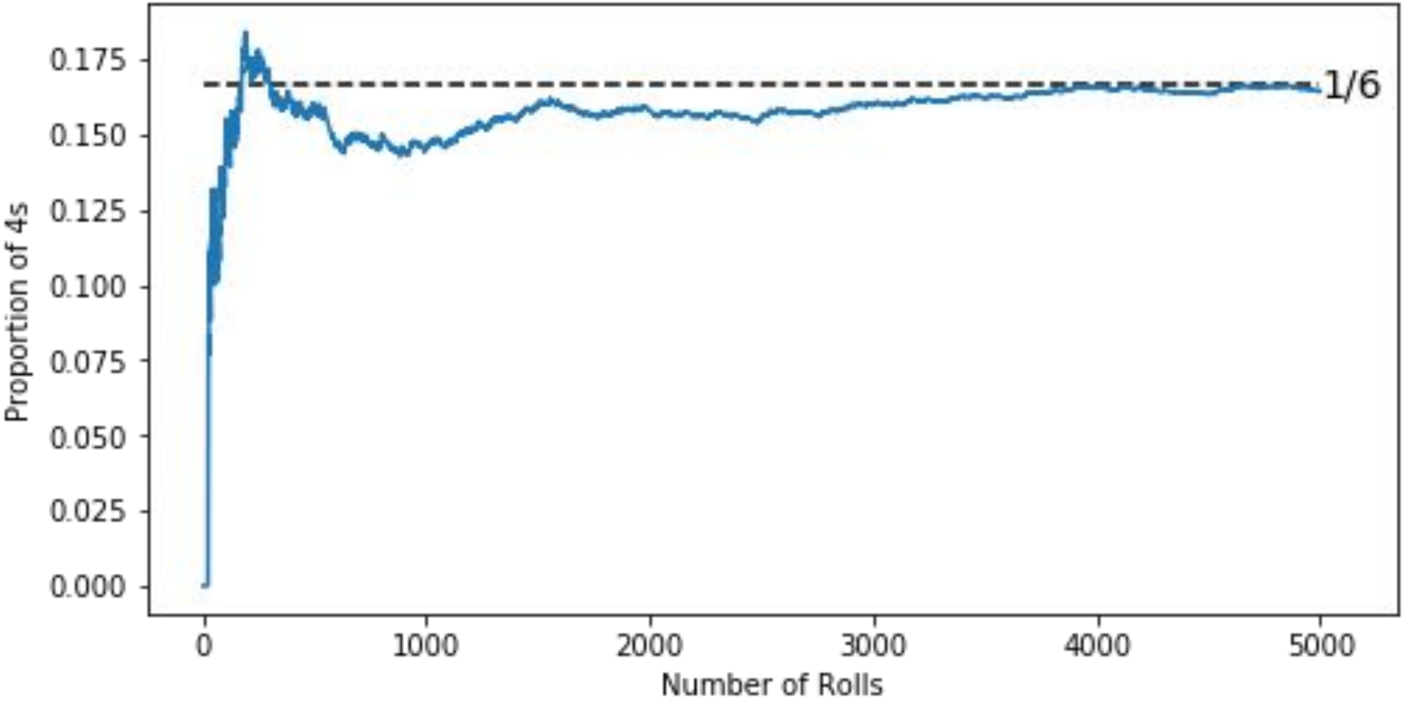
Probability as **Relative Frequency**

What do we mean when we say that the chance of rolling a 4 is 1/6?

Since there are 6 equally likely faces, if were to roll the die a large number of times, then it should land on each face approximately (probably not exactly) 1/6 of the time.

If we continued to roll the die over and over and tracked how often it landed on 4, the proportion of times that the die lands on 4 should approach 1/6.

Probability as **Relative Frequency**



Probability as **Relative Frequency**

The probability of event *A* is the proportion of times that *A* occurs in a infinite sequence (or very long run) of separate tries.

Law of Large Numbers: According to the 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 to the expected value as more trials are performed.

Probability Terminology

In probability theory, we talk about the probability of an event.

Notation: For event *A*, the probability of *A* is written as *P(A)*

Probabilities range from 0 to 1, with 0 representing *impossibility* and 1 representing *certainty*.

Probability Terminology

Experiment: A random process. Produces an outcome which we cannot know ahead of time.

Outcomes: The possible results from the experiment

Sample Space: The set of all possible outcomes

Probability Law: Assigns a probability to each possible event

Events: A collection of some number of outcomes

Conditional Probability

How does the likelihood of a particular outcome change, if we are given more information?

Eg. What is the probability that a person develops lung cancer?

What if we know that person smokes? How does the probability change?

Conditional Probability

For two events *A* and *B*, the **conditional probability of *A*, given that *B* has occurred** is *P(A / B)*

Probability of developing lung cancer:

P(lung cancer)

Probability of developing lung cancer, given that a person smokes:

P(lung cancer | smokes)

Conditional Probability

To find P(A|B), we need to realize that we are looking for the proportion of outcomes where both *A* and *B* occur, but only need to consider those where *B* has occurred.

We can calculate the conditional probability as

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

Conditional Probability

So if the 14% of the population smokes and 3.25% of the population both smokes and develops lung cancer, then

$$P(\text{cancer} \mid \text{smokes}) = P(\text{cancer and smokes}) / \mathrm{P}(\text{smokes})$$

$$= 0.0325 / 0.14$$

$$= 0.232$$

Probability as **Degree of Belief**

Thinking in terms of conditional probabilities, we can understand probability in a different way.

Probability quantifies how *certain* we are about a given hypothesis (event).

This interpretation of probability forms the basis for **Bayesian** statistics.

Bayes’ Theorem

Say that there is a disease that is carried by 0.1% of the population.

A test is developed which gives a positive result 98% of the time when a person has the disease.

It also gives a positive result 3% of the time when a person does not have the disease.

You get tested and the get a positive result. How worried should you be?

Bayes’ Theorem

Disease that is carried by 0.1% of the population:

$$P(\text{disease}) = 0.001$$

Positive result 98% of the time when a person has the disease:

$$P(\text{positive} \mid \text{disease}) = 0.98$$

Positive result 3% of the time when a person does not have the disease:

$$P(\text{positive} \mid \text{no disease}) = 0.03$$

Bayes' Theorem

You get tested and the get a positive result. How worried should you be?

Want to know: P(disease | positive)

Bayes' Theorem

$$\underset{\text{Posterior}}{\underset{\uparrow}{P(A|B)}} = \frac{P(B|A) \cdot \overset{\text{Prior}}{\overset{\searrow}{P(A)}}}{P(B)}$$

Bayes' Theorem

$$P(\text{disease}|\text{positive}) = \frac{P(\text{positive}|\text{disease}) \cdot P(\text{disease})}{P(\text{positive})}$$

$$= \frac{(0.98) \cdot (0.001)}{P(\text{positive})}$$

But, what is P(positive)?

Bayes' Theorem

But, what is P(positive)?

To answer this, we can use the fact that having the disease and not having the disease are *mutually exclusive.*

P(positive) =

P(positive and disease) + P(positive and no disease) =

P(disease)*P(positive|disease) +

P(no disease)*P(positive|no disease) =

(0.001)*(0.98) + (0.999)*(0.03) = 0.03095

Bayes' Theorem

$$P(\text{disease}|\text{positive}) = \frac{(0.98) \cdot (0.001)}{(0.03095)}$$

$$= 0.0317$$

So testing positive only increases your chances of having the disease by about 3%.

Conditional Probability

Independent and Dependent Events

Events *A* and *B* are **independent** if the occurrence of *B* in no way informs us about the probability of *A*. The two events do not “interfere” with each other.

That is, P(A|B) = P(A)

Equivalently, P(A and B) = P(A)*P(B)

Otherwise, we way that *A* and *B* are **dependent**.

Conditional Probability

Example: Flipping a coin twice

P(Second Coin Landing on Heads) = 0.5

P(Second Coin Landing on Heads | First Coin Landing on Heads) = 0.5

That is, knowing the outcome from the first flip gives us no additional information about the next one.

Conditional Probability

Why do we care?

When drawing a sample, we usually assume that the individuals are drawn *independently*.

When we’re looking at a dataset, we usually assume that our observations are independent of each other.