Difference between probability and statistics

Photo by Riho Kroll on Unsplash

As a famous mathematician Karl Pearson said:

> "Statistics is the grammar of science."

Statistics lies at the heart of practically all technological advancements and enjoys application in wide range of fields from biostatistics, finance, economics to machine learning and even more.

Having said that, we will not any book on Statistics which does not talk about the concepts of probability. In fact, they are closely associated branches of mathematics which are intertwined with each other.

In this article, we will understand what is the difference between the two. And, then we will conclude with the ubiquitous question that all of us might have faced at some point - which one to study first?

Probability:

Probability is a forward-looking concept.

It seeks to predict the next output given our knowledge of the distribution of the data generating model which is a random process i.e. $P(Y|\theta)$

where:

Y: outcome

θ: parameter that defines the random process, where random process which could be flipping a coin, rolling a die etc.

There are two school of thoughts to interpret probability:

1) Frequentist: It concerns around calculating the relative frequency of occurrence of a certain outcome upon repeating the experiment multiple times
   - Let's take an example of a fair coin toss. Tossing a coin 10 times might show 7 'Heads' which converts into a probability of 0.7, however, if we repeat this experiment long enough, then the probability of 'Heads' starts converging to 0.5.
2) Bayesian: It primarily talks about
   a. the degree of belief captured in the term known as 'prior'
   b. updated information from the experimental data, incorporated in 'likelihood' denoted by L(θ|Y).

   The probability is then updated by taking product of prior with the likelihood

$$P(Y/\theta) \propto P(Y) * L(\theta/Y)$$

Let's understand more about likelihood: As real-life processes are highly complex that we often don't have a fair idea of the distribution of the stochastic process, we need to have a better estimate of θ in order to be able to better predict the P(Y/ θ).

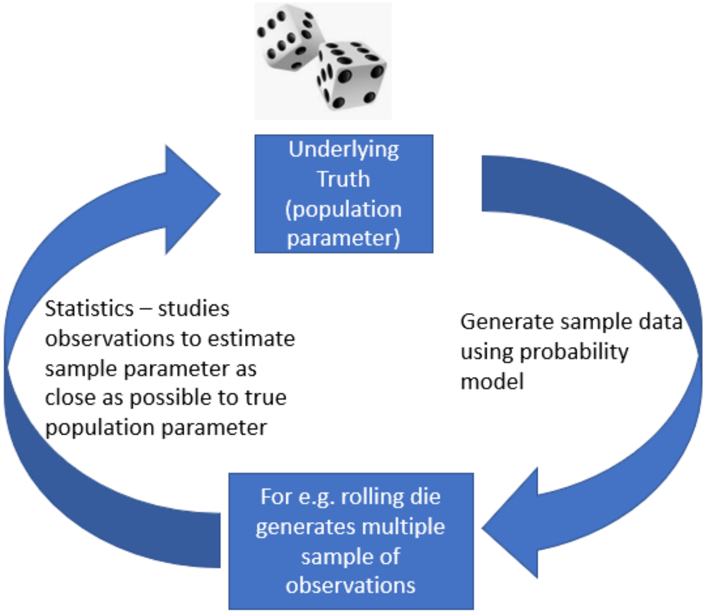
Simply put, we need to estimate the θ that best describes the distribution from which given data is generated. This is where Statistics comes into picture – to calculate the likelihood function.

Let's learn more about Statistics

Statistics:

1) Statistics is backward looking i.e. it extracts the patterns from the underlying random process to come up with a possible quantitative model that explains the data behavior.
2) Note that we do not necessarily know all the characteristics that define the underlying random process.
3) Statistical model starts with a finite possible values θ (called as model space Θ) along with the observed data to infer which θ would have generated that data.
   a. Where θ could be any parameter that defines the probability distribution like mean or variance.
4) Statistical modelling also helps us to quantify the error around the inferred distribution or parameter.
5) Estimated parameter $\hat{\theta}$ gets closer and closer to the true θ as more data is collected from the same experiment. That is, the error around the estimates shrinks and we become more confident in our estimates. This is called as confidence interval.
6) Now, you will be thinking from where does this error arise and how much data is enough to infer the true parameter.
7) Well, that's because we only get to study a sample of data and not the entire population for various reasons like data availability in first place, cost of accessing the data etc. If with repeated

experiments we can get our sample space so close to the population, we could reduce the error of our sample parameter estimates to 0 and arrive at true parameter.


Underlying Truth (population parameter)
Generate sample data using probability model
For e.g. rolling die generates multiple sample of observations
Statistics – studies observations to estimate sample parameter as close as possible to true population parameter

Difference between Probability and Statistics – explained via examples

Example 1:

Probability: For a drug trial, the entire population is divided into two groups in 1:9 ratio – A and B. P(A) = 70% and P(B) = 50%, where P = probability of successful drug effect. What is the probability of successful treatment of a random patient from the population?

Statistics: Now, if we do not know the true population composition and observe the data for 700 successful patients among 1000 randomly selected patients from the population. What can we infer about the population distribution?

Example 2:

Probability: If we flip an unbiased coin 3 times successively, what is the probability that we get HTH? That is, given we know the parameters what is the probability of observing a particular occurrence HTH?

Statistics: If we observe HTH outcome of 3 successive coin flips, what is the estimate of parameter p of the binomial distribution, where p = probability of getting heads?

Example 3:

Probability: From a pack of cards comprised of 4 suits of Spades, Hearts, Diamonds, and Clubs each, what is the probability of drawing 4 cards - 2 hearts and 2 clubs from the pack?

Statistics: If we don't know the playing cards composition and observe a random sample of 4 cards produced 1 heart, 2 spades and 2 clubs. What can we infer about the total distribution proportion in the card?

What comes first: probability or statistics:

Statistics which is the core of machine learning, works on collecting lot of data to be able to generate insights and drive decision making. Statistics includes but is not limited to the study of probability models only.

There are multiple opinions and discussions mentioned here, however I would like to conclude that it is an iterative process to understand the foundations of probability enough to understand the central dogma of statistics.

In the process, when we start understanding the concepts behind statistics and probability models, we are no more confined by pedagogy of what to be studied first.

References:

https://stats.stackexchange.com/questions/2641/what-is-the-difference-between-likelihood-and-probability

https://johnkerl.org/doc/prbstat/prbstat.html