Probability and Measure

let's suppose that you are a student of mathematics and have gone as far as learning integral calculus. Suppose you want to broaden your horizon by learning some probability and statistics.

You will likely find that the math in probability and statistics is not too difficult, once you have learned the terminology. However, do not expect statistical inference to be soundly grounded on axioms and proofs. You cannot prove the validity of statistical methods. Think of statistics as an art form that extends common sense and is guided by the mathematics of probability theory. We wind up with a convincing method of extending our ability to make decisions with limited information.

Probability

To begin with, the idea of probability is understood by most people as an idealization of ignorance. We easily know that the probability of tossing a coin and getting heads is ½. In this thought experiment, we assume that we have an ideal coin (which doesn't exist). If we toss 100 heads in a row, we suspect that the coin is not near ideal.

Here, we take the idea of probability as an undefined concept. Starting with undefined concepts is a typical logical trick which is used throughout mathematics, like the undefined points and lines in geometry. Just because a concept is undefined, doesn't mean that we don't have a common understanding of what it means.

We will be examining cases of discrete probability such as tossing a coin, and cases of continuous probability such as the final resting angle of a spinning needle, or dial.

Measure Theory

Probability involves integration. An advanced study of integration involves measure theory. We won't be going there, but we will use some of its ideas. let's look at Lebesgue measure on the real line without formal definitions.

Lebesgue measure gives the size of subsets of the real line. The measure of an interval is its length. The measure of a set composed of many disjoint intervals is the sum of their lengths. The measure of a single point is zero. Lebesgue measure comes into the theory of integration because it gives us a language for breaking the domain of integration into little pieces. Incidentally, there are subsets of the real line that do not have Lebesgue measure. There is a significant rabbit hole here involving non-measurable sets and we will leave it alone.

You can begin to imagine how Lebesgue measure applies to probability by noting that there are cases where the probability of landing in an interval is proportional to the length of the interval.

Sometimes, we will use slightly different notation for integration.

Instead of writing $\int_a^b f(x)\,dx$ we will write $\int_{(a,b)} f(x)\,dx$.

Suppose we have a set that is composed of disjoint open intervals like $A = \bigcup_{i=1}^{n} (a_i, b_i)$. Instead of writing $\sum_{i=1}^{n} \int_{a_i}^{b_i} f(x)\,dx$ we will write $\int_A f(x)\,dx$.

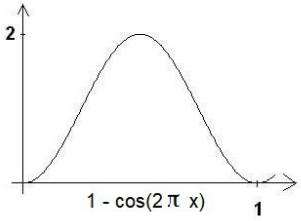
In other words, we will be writing the domain of integration below the integration symbol.

Spinner Example

Consider an ideal spinning dial (or needle) that will stop at any angle randomly. If I paint a $\pi/4$ radian sector on the face of the spinner, we know that the probability of landing in the sector is 1/8.

We will consider the spinner further by sketching the angles from 0 to 2π on the x-axis of a graph over the interval [0 , 1] (scale 0 to 2π to the unit interval) . We express uniform probability of the angles by stating that the probability of landing in a half-open interval (not crossing the $0 - 2\pi$ boundary) is given by its length $P(a,b) = b - a$.

We can imagine modifying the spinner so that it has continuously varying friction, with more friction near π radians, so that it is more likely to stop near π. Its probability density (or distribution) might be expressed as $p(x)=1-\cos(2\pi x)$. Then $P(a,b)=\int_{(a,b)} p(x)\,dx$ is larger for angle intervals near $x=0.5$, which corresponds to angles near π. This is described as more weight near 0.5.



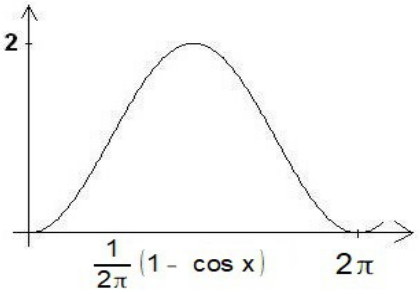
Note: $P(0,1)=\int_{(0,1)}\left(1-\cos(2\pi x)\right)dx=1$

Terminology

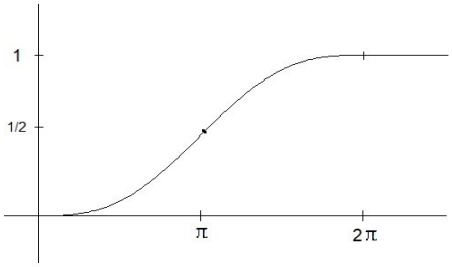
We will use the spinner example to exhibit some terminology. Individual points in the unit interval (or angular positions of the pointer) form the set of outcomes (or sample space). Events are intervals (or collections of intervals). Probabilities are real values assigned to events (via an integral). Probability values are between zero and one. We interpret the probability values zero and one as impossible or certain. For the spinner example, the probability of landing on a single point (a degenerate interval) is zero. Probability is additive in the sense that the probability of landing in a collection of preselected disjoint intervals is the sum of the probabilities of landing in each interval. We express the probability of landing in a set of intervals A, as an integral of the density function $P(A)=\int_A p(x)\,dx$.

An important feature of the density function is $\int_{\text{all outcomes}} p(x)\,dx=1$.

In the spinner example, we can just as easily represent the possible angles on the interval [0, 2π]. In this case, we would express the density function for uniform probability as the constant function $p(x)=\pi/2$. If we wanted to represent our previous bias towards π, we would express the density (ie. re-scaling) as $p(x)=\frac{1}{2\pi}(1-\cos x)$



In many cases, we use the cumulative distribution function which is given by $F(x)=P(-\infty,x]=\int_{(-\infty,x]} p(t)\,dt$. In our last example $F(x)=\frac{1}{2\pi}(x-\sin x)$ (with F(x) = 0 for x < 0 and F(x) = 1 for $x>2\pi$).



When the cumulative distribution function is available, you can do probability calculations without integration. For example:

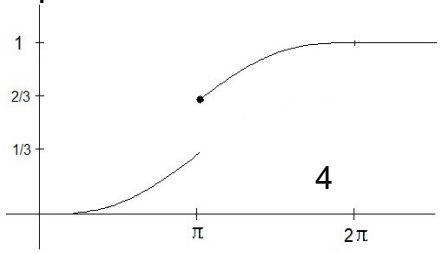
$$P(a,b] = F(b)-F(a)$$

In many cases, the derivative of the cumulative distribution function is the density function: $\frac{d}{dx}F(x) = p(x)$

Discrete Probability

The continuous probability density for the spinner doesn't describe discrete processes like coin flipping well. let's modify the spinner model with a drop of viscus syrup at the π radian point so that it has a probability of 1/3 of stopping at π and a uniform probability of 2/3 of stopping anywhere else. Then probability calculations involve integration over intervals with the possibility of adding 1/3 if π is included in the interval. There are difficulties with describing the density at π with a density function, which should represent an infinite density at π . One approach (which we won't follow) would be to allow density functions to include generalized functions like the delta function. Then the density function for the above case could be expressed on the interval $[0\ ,\ 2\pi]$ as $p(x) = \frac{2}{3}\left(\frac{1}{2\pi}\,(1-\cos x)\right) + \frac{1}{3}\,\delta_\pi$. In any case, we think of discrete probability of an outcome as a point mass (or point measure). We usually do not use generalized functions. We simply numerically add the discrete probability when necessary.

We will not use a density function here, but we can represent this with a cumulative distribution function. The discrete probability at π is represented by a discontinuity in the cumulative distribution function without the annoying infinity for its density at π .



We can express probabilities for open intervals as $P(a,b) = F(b) - F(a)$.

Discrete probability allows us to express probability for an event that contains a single outcome (or sample point). Here, $P(\{\pi\}) = \lim_{\epsilon \to 0} P(\pi - \epsilon\ ,\ \pi + \epsilon)$.

Coin Tossing
Consider sequentially tossing three fair coins. There are eight possible outcomes. Under our assumption of the fairness of the coins, all eight outcomes are equally probable. For any event (subset of the eight outcomes), we calculate the probability by counting the outcomes in the set:

The probability of tossing three heads is 1/8.
The probability of tossing exactly two heads is 3/8.

Note that we can assign a probability to any event in this case. In the context of continuous probability, there are events that cannot be given a probability. Typically, these events are pathological non-measurable sets that cannot be expressed as a countable union of intervals. We won't be going there.

Combinatorics
There are many practical applications based on the discrete probabilities of rolling dice or selecting objects from urns. Typically, we count the number of possible outcomes and then assume all outcomes are equally probable. So, the critical technique involves the counting of various patterns. The main counting formulas are illustrated as:

The number of sequences resulting from rolling n dice with k-sides: n^k .

The number of ways of rearranging a list of n things: $n!$

The number of k element subsets of n things $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$

Combinatorics can get tricky!

Random Variables
To analyze a real-world problem, you will need to choose idealizations that are mathematically tractable. In this context, the first step is to identify a space of outcomes and plausible probabilities for sets of outcomes. This process is called selecting a random variable.

A random variable is a function from the space of outcomes to the real line, and a cumulative distribution function on the real line that is compatible with the assumed probabilities in the real world.

The formalism of a random variable is a little confusing because the domain is in the physical world, and the range is in the abstract mathematical world. However, this is the type of thinking you use whenever you do applied mathematics.

In the spinner example, the physical spinner is in the real world. The random variable lies between 0 and 2π . An evaluation of the random variable x might look like: $x(\text{six o'clock}) = \pi$

The distribution is chosen according to context and might be uniform or $\frac{1}{2\pi}(1 - \cos x)$.

In statistical terminology you might hear:
"Let x be a normally distributed (Gaussian) random variable."
"Let x be a uniformly distributed random variable."

The formalism of random variables for discreet processes like coin tossing can be a little awkward. Maybe heads and tails should be mapped to {0,1}, or {-1, 1}. In cases like this, it doesn't add to understanding, by arbitrarily choosing two points on the real line with equal probability. In this case it is easier to think of heads and tails and possibly the mathematical set {H,T} without embedding it in the real line. You shouldn't think too much here.

Expected Value

We all know how to take the average of n numbers as: $\bar{x} = \sum_i \left(x_i \frac{1}{n} \right)$. We also know about taking a weighted average when calculating a class grade where the scores from homework, midterms, and final exam have different weights:
$\bar{x} = \sum_i (x_i \alpha_i)$ where $\sum_i \alpha_i = 1$. The analogous calculation in the context of probability is:
$\text{mean} = \bar{x} = E(x) = \int x\, p(x)\, dx$ the expected value of x.

(we usually omit denoting the domain of integration when it is the entire sample space.)

This convention continues with
$E(x^2) = \int x^2\, p(x)\, dx$ the expected value of x^2

$k^{\text{th}}\text{moment} = E(x^k) = \int x^k\, p(x)\, dx$ the expected value of x^k

The expressions x, x^2 and x^k are all functions with the range of the random variable as their domain. In general $E(f) = \int f(x)p(x)dx$ means the expected value of f, which is synonymous with the average value of f.

Functions of a random variable can create their own random variables via composition with the original random variable. The distribution of this new composed random variable must be compatible with the original random variable and can be hard to calculate.

Gaussian Density and Random Walk

Consider a grid of equally spaced points along the x axis starting at zero. A bug starts at zero and repeatedly jumps to the left or right depending on the flip of a fair coin. We can calculate the probability of being at a grid point after n jumps. Most of the weight of the probability will be near zero, and this discrete probability looks like a familiar bell shaped curve. We judiciously choose the grid spacing to be $\Delta x = \frac{1}{\sqrt{n}}$. After some combinatorial calculations we could see that the probability of being at $x = \frac{2k-n}{\sqrt{n}}$ after n jumps is given by

$$P(\{x\}) = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{|k|}$$

The mean of this probability is zero because of symmetry. Because of the careful choice of grid spacing, $E(x^2) = 1$. For large n, $P(\{x\})$ is similar in appearance to the standard bell shaped Gaussian density given by :

$$p(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

In fact, for any interval (a,b] we have $\sum_{x\in(a,b]} P(\{x\}) \underset{n\to\infty}{\to} \int_{(a,b]} p(x)dx$.

This illustrates why the Gaussian density is used so much. The central limit theorem tells us that we always get this behavior for random variables built from adding repeated outcomes from independent identically distributed random variables (with mean zero).

Independence

In the 3-coin tossing example, we unconsciously used the idea of independent tosses. We assumed that we had a fair coin and the tosses don't influence each other. We then judge that all eight outcomes are equally probable and use this to calculate the probability of any event. If the coins were strung together on a rubber band so that they influenced each other, we would use another model.

Mathematically, we express independence as a new probability function on pairs of events as a simple product rule: P(A,B) = P(A) P(B). For example, if we flip a coin and spin a simple spinner (uniform density):

$$P(\{H\},\{\theta:\ \pi/4 < \theta \le \pi/2\}) = \left(\frac{1}{2}\right)\left(\frac{1}{8}\right)$$

Simple intuition sometimes subverts our use of independence. If you avoid choosing a previous winning lottery number, you should know of a mechanism that reduces repetition (like selection without replacement). However, lottery managers work very hard at making winning numbers independent.

Conditional Probability

Conditional probability is used to model an increase in information. This is usually expressed as a reduction in the set of possible outcomes. In the 3-coin example, the probability of tossing exactly two heads is 3/8. However, the probability of tossing exactly two heads conditioned on the first toss being heads is 2/4. The calculation has a set theory flavor on the probability of events. Let X be the set of all 3-coin sequences. Let A be the set of outcomes with exactly two heads. Let B be the set of all 3-coin sequences that begin with H. Then the conditional probability is expressed as

$\frac{P(A \cap B)}{P(B)} = \frac{2/8}{4/8} = \frac{2}{4}$ note: $A \cap B$ is the set of all sequences with exactly two heads that begin with heads.

Ambiguous Distributions

Practical problems involving probability calculations are based on an appropriate probability distribution assumed by the random variable. Sometimes the distribution that should be used appears obvious when it isn't.

There are famous paradoxes where the context of the question does not give guidence on the underlying distribution. Bertrand's paradox involves finding the average length of the cord of a circle. The wine and water paradox involves finding the average proportions in a wine and water mixture chosen at random.

Always be suspicious when you hear the phrase "choose at random"!, with no specified distribution.

Statistics

Probability theory is part of mathematics and is soundly grounded in axiomatic and logical methods. Now consider a common sense activity where you see a coin flipped 100 times and see it come up heads every time. You would doubt that it is a fair coin. It is possible that it is a fair coin with an unexpected outcome.

However, it makes sense not to trust the coin even though it appears to be a normal coin. You can argue that a fair coin would give this result with probability $1 / 2^{100}$, and thereby quantify your doubt. This doesn't prove anything in the mathematical sense. Note that a full run of heads is equally probable to any other sequence, even though it is subjectively special. But, you have quantified and sharpened your common sense with a probability calculation. The first 100 binary digits of π is not random, but it certainly looks that way. Since most of us don't know what they are, we might as well treat them as random. But, not in the context of measuring circles.

Statistical inference is an art and science where probability calculations inform a judgment. Typically, there is a subjective choice of model expressed as a set of outcomes and a probability distribution. In the coin tossing example, our experience of coin properties leads us to our standard model of coins, and choice of the standard binary random variable. The choice of random variable and subsequent calculations partly depends on tradition, and what seems convincing.

Most practical applications involve sampling from a larger population. In this case, the variable involves judgment on the distribution of the population of samples. Some random variables are universally convincing, such as the fairness of a coin or the repeated independence that occurs in a random walk justifying use of the Gaussian distribution. Sometimes the choice of model can be clearly wrong such as a naive random walk model for stock prices, where individual transactions are neither identically distributed nor independent. Financial statisticians persist with this model because there is nothing better.

Ultimately, statistical inference is based on belief in a suitable ideal random variable in the same way a physicist models the flight of a ball as an ideal point mass (perhaps in a vacuum). However, choosing the right random variable to express our ignorance in a given situation is more subtle than the choice of a Newtonian physics abstraction.

Ignorance and Information from Sampling

Suppose we wish to make a statistical argument via sampling. We view the world through our previous knowledge and expectations. In many cases, we want to shield our analysis from our bias. We might mistrust our expectations and follow a procedure similar in spirit to doing a check-sum in arithmetic. Sometimes we wish to argue that an analyses is not just personal opinion. In either case, we need to argue from an acceptably random sample. In other words, we improve the validity of our argument by increasing the ignorance of our sample.

We often cannot create a procedure for getting an unbiased sample. For example, the willingness of someone to answer a questionnaire biases the sample. We may tinker with the analysis of the sample by using information

about the relationship between the sampling method and the population. This is an art and requires subjective judgment. Sometimes the population being sampled is hypothetical or in the future (such as political polling). Sometimes sampling effects the population (such as push polling). An important part of a statistical argument is the credibility of the sampling method and its treatment.