Recitation 3: Key Concepts in Probability

Christina Su

October 18, 2018

This recitation is intended to provide an overview of key topics in probability theory needed to understand the methods of statistical inference taught in this course. Much (or all) of the material may be review, so feel free to read as much or as little as you choose.

1 Introduction to Probability

1.1 Definition

"Probable" and "likely" are concepts that come up frequently in daily life, whether talking about the probability of having rain the next day or the likelihood of getting food poisoning from eating Chipotle. Fundamentally, *probability* is defined as the likelihood that a given event or outcome will occur. (As in lecture, we will use this intuitive definition rather than giving a precise mathematical formulation.) Specifically, we let $P(A)$ denote the probability that event A occurs. Examples of events of interest might be drawing a pair of aces from a deck of cards, correctly answering a series of True/False questions, or finding a job in your field of interest within 3 months.

1.2 Interpretations

We now have a general framework for thinking about probability, but our definition ("the likelihood that a given event or outcome will occur") is still somewhat vague and open to interpretation. Indeed, there are two leading schools of thought regarding interpretations of probability: frequentism and Bayesianism. To illustrate the difference in philosophy, we'll first consider how each school would interpret the probability that a given day in Pasadena will be sunny. It is important to note that these are differences in interpretation, not computation; all of the rules of probability that we will discuss below, including Bayes' theorem, hold true regardless of which school of thought you may subscribe to.

1.2.1 Frequentism

In the frequentist view, probabilities are objective claims about the physical state of the universe. More specifically, probabilities measure the long-run frequency of

the outcome of interest in many repeated experiments, which converges to the true frequency as the number of trials approaches infinity. To think about the probability of a sunny day, a frequentist would ask the following question: over the next hundred or thousand or million days, what is the frequency of sunny days?

1.2.2 Bayesianism

Bayesian probabilities represent more subjective measures of the degree of belief in a particular claim. In our example, a Bayesian would ask the following question: how plausible is the claim that it will be sunny today?

1.2.3 Biological Application

Let's now give an example of how these interpretations might apply in a biological context, rather than mundane small talk. Imagine that we want to study a gene called *SMART* that is associated with intelligence (yes, purely fictional). We want to look at the mean expression of *SMART*, denoted S, in Caltech students, our sample for highly intelligent individuals. Specifically, we want to know the probability $P(S > 10)$. A frequentist will approach this question as follows: if I repeat this study many (infinitely) more times (or in many more "universes"), how frequently will I find $S > 10$? A Bayesian will think about finding this value as follows: given the values I measured, how plausible do I think it is that $S > 10$?

1.3 Properties

Until now, we have only discussed the conceptual underpinnings of probability. Now that we have a better understanding of what we are trying to quantify, how do we quantify it? We'll start with some fundamental properties that probabilities must satisfy.

1. Conceptually, the extremes of probability are "impossible" and "certain." Mathematically, those extremes are defined as 0 and 1. If A has no chance of occurring, then $P(A) = 0$. If B is certain to happen, then $P(B) = 1$. Events that are possible but uncertain fall somewhere in between, with a higher probability denoting a higher likelihood of that outcome.
2. In general, something always has to happen. We denote the *complement* of A, or the set of all outcomes not represented in A ("not A"), as $\bar{A}$ or A^C. If I send an email to someone, he or she will either reply (A) or not reply ($\bar{A}$). Therefore, the union of those two events (the set of outcomes where at least one of the

two events occurs), denoted $A \cup \bar{A}$, satisfies

$$P\left(A \cup \bar{A}\right) = 1$$

We then have

$$P(A) + P\left(\bar{A}\right) = 1$$

(Most people will make that jump intuitively, but we'll later see more formally why it holds.)

2 Calculation of Probabilities

We now have some fundamental definitions of probabilities. However, the probabilities of impossible or certain events aren't particularly interesting, and we often want to quantify more complex scenarios by relating the probabilities of multiple events. In this section, we will start with the simplest approach to computing probability and then develop some rules for combining probabilities of multiple events.

2.1 Counting Approach

If asked about the probability of rolling a 1 with a fair die, most people will respond "$1/6$" without hesitation. How was that number obtained? Often, people are going through a counting approach, a basic procedure for calculating probability, so quickly they hardly need to think about exactly what they are doing. Specifically, we ask the following: how many possible outcomes are there, and in how many of those cases does our event of interest happen? Mathematically, we compute $P(A)$ as

$$P(A) = \frac{\text{number of events in which } A \text{ happens}}{\text{total number of events}}$$

In our example of rolling a 1 with a fair die, there are 6 possible outcomes, 1 of which is the desired outcome. Therefore, as you all already calculated instantly, the probability is $1/6$.

When in doubt, this process of enumerating the number of total and desired outcomes is a good starting point. Indeed, while this approach may seem simple, it can be combined with combinatorics to compute probabilities of quite complex events. Let's consider the Sorting Hat's job of sorting students into the 4 houses of Hogwarts. We want to know the probability that exactly 2 of the first 3 students are sorted into Gryffindor. Each student can get sorted into 4 houses, so there are $4 \cdot 4 \cdot 4 = 64$ possible outcomes. There are then 9 combinations that yield the desired outcome; the student not in Gryffindor can be in any of 3 houses, and that student can be first, second, or third. Thus, the desired probability is $9/64$. (We'll get to the unstated assumption of this calculation shortly.)

2.1.1 Combinatorics

The above example shows how combinatorics can be a useful tool in enumerating outcomes. We'll briefly state general formulas for the most common applications, but we won't spend too much time on the details. We often want to consider the total number of samples of size k that can be drawn from a population of size n. These samples may be either permutations, where the ordering of the items in the sample matters, or combinations, where the ordering does not matter. In addition, the samples may be drawn with or without replacement; that is, after a given item is drawn, it can be put back into the population for the next draw (with replacement) or removed permanently (without replacement). The total number of permutations drawn with replacement is n^k. For instance, in the example above, we were drawing a sample of size 3 (number of students) from a population of size 4 (number of houses). The number of permutations P drawn without replacement is

$$P = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

The number of combinations C drawn without replacement can be computed from the above formula by dividing by $k!$, or the number of possible orderings of the sample (that no longer need to be distinguished). This quantity, which is often referred to as "n choose k," is

$$C = \binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

2.1.2 Caveat

So, why doesn't this approach of enumerating outcomes work in all cases? First, in more complex scenarios, you may not want to (or be able to) enumerate the outcomes. For instance, when considering the probability of a sunny day, it would be essentially impossible to describe all possible configurations of atmospheric and geologic conditions. Second, counting relies on the assumption that each of the possible outcomes is equally likely. Going back to the Sorting Hat example, if the Sorting Hat places students in Gryffindor more often than in any other house, we can no longer just count events, as we will not be properly weighting the outcomes by their varying probabilities. Therefore, we will now turn to developing rules of probability that hold in the general case.

2.2 Union of Events

Formally, the *union* of events $E_1, E_2, \ldots, E_n$ is defined as the set of outcomes that include any of $E_1, \ldots, E_n$. Thus, the probability $P(E_1 \cup E_2 \cup \ldots \cup E_n)$ represents

the probability that at least one of the n events of interest happened. For now, we'll stick to looking at the union of 2 events, but this analysis can be extended to an arbitrary number of events.

We'll start with an example. Imagine that we would like to study the wellness habits of Caltech students. Specifically, we are interested in looking at patterns of diet and exercise. Let $P(A)$ denote the proportion of students who eat at least 5 servings of fruits and vegetables per day and $P(B)$ the proportion of students who exercise at least 2 hours per week. First, we want to know the probability that a given student either eats well or exercises regularly, or $P(A \cup B)$. We can start by computing $P(A) + P(B)$, but students who satisfy both A and B will be counted twice. Therefore, we correct our formula by subtracting the probability of the intersection of A and B, or the probability that both events occur; we denote the probability of the intersection as $P(A, B)$ or $P(A \cap B)$. Our formula for the probability of the union of two events is

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

2.2.1 Mutually Exclusive Events

Let's now say we want to look at the extremes of exercise levels, or the probability that a student either does not exercise at all (A) or exercises more than 20 hours per week (B). In this case, A and B are mutually exclusive events; by definition, they cannot occur simultaneously, or $P(A, B) = 0$. Therefore, the formula for the probability of their union simplifies to

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

This *sum rule* applies when analyzing the union of mutually exclusive events.

2.3 Intersection of Events

We now want to compute the proportion of Caltech students that both eat well (A) and exercise regularly (B). We are therefore considering the *intersection* of events, or the set of outcomes in which all of the events are included. Therefore, the probability $P(E_1, E_2, \ldots, E_n)$ represents the probability that all n events of interest occur. How would we compute $P(A, B)$? We can start by looking at the probability that one of the events occurs, or $P(A)$. We can then define $P(B)$ in relation to A. Specifically, we define the *conditional probability* $P(B|A)$, or the probability that B occurs given that event A occurred. However, we can also reverse the analysis and start with the assumption that B has occurred. Thus, the intersection of A and B can be written equivalently as

$$P(A, B) = P(B|A)\,P(A) = P(A|B)\,P(B)$$

2.3.1 Independent Events

Let's suppose that we know that Caltech students who eat well have the same patterns of exercise as those who do not eat well. That is, the knowledge that A occurred does not tell us anything about $P(B)$. In this case, we say that A and B are *independent*, or $P(B|A) = P(B)$. Therefore, our formula for the intersection of events reduces to

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

This *product rule* applies to the intersection of independent events.

2.4 Conditional Probabilities and Bayes' Theorem

We have already introduced the idea of conditional probabilities above. However, the formula for the intersection of two events can be rearranged to give *Bayes' theorem*, which states

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

We can condition each probability on our prior information I. We will also replace A with H, representing our hypothesis, and B with D, representing our data. We now have

$$P(H|D,I) = \frac{P(D|H,I)\,P(H|I)}{P(D|I)}$$

This form tells us how to use our data in conjunction with prior knowledge to quantify our confidence in a given hypothesis. We can rewrite the equation in words as

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

In the coming weeks, we will delve much more deeply into defining likelihoods and priors and using them to compute posterior probabilities. For now, we will show some concrete examples of how to use Bayes' theorem to learn from our observations.

Example: Consider a multiple-choice test. For every question, a student has a probability p of knowing the answer. Otherwise, the student guesses, choosing at random from the m choices. The instructor is interested in finding out whether students actually know the material. Given that a student answered a question correctly, what is the probability that he or she knew the answer?

Solution: Let K denote the event that the student knew the answer and C the event

that he or she answered correctly. Using Bayes' theorem, we have

$$P(K|C) = \frac{P(C|K)\,P(K)}{P(C)}$$

We know that a student answers correctly when he or she knows the answer, or $P(C|K) = 1$. We also defined $P(K) = p$. Therefore, we need only calculate $P(C)$. A student can be correct from knowing the answer or from guessing it; since these are mutually exclusive events, we can apply the sum rule and our knowledge of conditional probabilities to find $P(C)$. Letting G denote the event that the student guessed, we have

$$\begin{aligned} P(C) &= P(C|K)\,P(K) + P(C|G)\,P(G) \\ &= 1 \cdot p + \frac{1}{m} \cdot (1-p) \\ &= p + \frac{1-p}{m} \end{aligned}$$

Substituting into Bayes' theorem, we find

$$\begin{aligned} P(K|C) &= \frac{1 \cdot p}{p + \frac{1-p}{m}} \\ &= \frac{p}{p + \frac{1-p}{m}} \end{aligned}$$

Example: As a graduate student in biology, I was warned early on that working with animals (particularly primates) is a surefire way to have a long PhD. Let's say I disregard that advice entirely and decide to start a project involving construction of a new mouse strain. Specifically, I want to test whether *SMART* knockdown reduces performance on tests of spatial learning. I design a construct expressing a nonfunctional form of SMART and a protein that makes mice spotted instead of solid in color (as a marker for my construct). This allele, denoted S, is dominant over the normal allele, s. That is, mice with an SS or Ss genotype will appear spotted, and mice with the ss genotype will be solid. I want to identify a mouse with an SS genotype that I can use for breeding. I have a candidate mouse in mind. It is spotted, and both of its parents are spotted. However, one of its siblings is solid. What is the probability that my candidate has the SS genotype?

Solution: The solid sibling must have genotype ss; therefore, it inherited one s allele from each parent. We then know that the parents both have genotype Ss. We can use this information as well as the observation that my candidate is spotted. From the definition of conditional probability, we can compute the probability that the candidate mouse has the SS genotype as

$$P(SS|\text{spotted, parents } Ss) = \frac{P(SS,\ \text{spotted, parents } Ss)}{P(\text{spotted, parents } Ss)}$$

Two parents with an Ss genotype will have offspring of the following genotypes: SS, Ss, sS, and ss. Thus, the probability of having an SS genotype is $1/4$, and the probability of having an Ss genotype is $1/2$. Substituting into the expression above, we find

$$\begin{aligned} P\left(SS|\text{spotted, parents } Ss\right) &= \frac{1/4}{1/4+1/2} \\ &= \frac{1}{3} \end{aligned}$$

Example: I'm not very happy with that probability, but I'm in a rush and continue with my experiments nonetheless. I breed my spotted mouse with a solid mouse. Given that all of its n offspring are spotted, what is the probability that my mouse really does have the SS genotype?

Solution: We start by writing Bayes' theorem as

$$P\left(SS|n \text{ spotted offspring}\right) = \frac{P\left(n \text{ spotted offspring}|SS\right)P\left(SS\right)}{P\left(n \text{ spotted offspring}\right)}$$

We already know every term in the numerator; all offspring of an SS mouse will inherit an S allele and therefore be spotted, and we calculated $P\left(SS\right)$ above. I omitted conditioning on I for convenience of notation, but it would be more appropriate to think of $P\left(SS\right)$ as $P\left(SS|I\right)$, or our prior belief in the hypothesis given our existing state of knowledge from above. We now need to compute the denominator. As above, we will consider the two possible ways to get the event of n spotted offspring and add their probabilities, finding

$$\begin{aligned} P\left(n \text{ spotted offspring}\right) &= P\left(n \text{ spotted offspring}|SS\right)P\left(SS\right) \\ &\quad + P\left(n \text{ spotted offspring}|Ss\right)P\left(Ss\right) \\ &= 1\cdot\frac{1}{3}+\left(\frac{1}{2}\right)^n\cdot\frac{2}{3} \\ &= \frac{1}{3}\left(1+\frac{1}{2^{n-1}}\right) \end{aligned}$$

Substituting into Bayes' theorem yields the desired probability as

$$\begin{aligned} P\left(SS|n \text{ spotted offspring}\right) &= \frac{1/3}{\frac{1}{3}\left(1+\frac{1}{2^{n-1}}\right)} \\ &= \frac{1}{1+\frac{1}{2^{n-1}}} \\ &= \frac{2^{n-1}}{2^{n-1}+1} \end{aligned}$$

As I observe more and more spotted offspring, or data which have a high likelihood if my hypothesis is true, the probability that my hypothesis holds increases. In the Bayesian sense, I become increasingly more confident that my mouse is indeed *SS*.

Example: An unknown number of years later, I finish my experiments and write up a paper, which I need in order to graduate. Several months later, I hear back from the editor, who says that my paper is one of three papers under consideration that are all of equal quality. However, the journal can only publish two of them, so one will be rejected at random. Disheartened, I ask if I can at least know one of the papers that will be published from the other submissions, since I already know that one of those two must be published. The editor asks if I really want to hear it, since my probability of being rejected will go from $1/3$ to $1/2$, as one of only two remaining candidates. Is it really true that my hopes of graduation will fall if I ask for this information?

Solution: Let's call the other two papers A and B. Suppose I confirm that I want the information, and the editor tells me that A will be published. (We only need to consider this case, as we can interchange A and B without loss of generality.) To calculate my new probability of being rejected, we'll write out Bayes' theorem, which states

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

If my paper is rejected, the editor can tell me A or B; we'll assume that this choice is made at random, so $P(A|\text{rejected}) = 1/2$. Since the rejection is decided randomly, we know that $P(\text{rejected}) = 1/3$. To calculate $P(A)$, we need to consider two possibilities: B is rejected, or my paper is rejected. We compute this value as

$$\begin{aligned} P(A) &= P(A|B \text{ rejected})\,P(B \text{ rejected}) + P(A|\text{rejected})\,P(\text{rejected}) \\ &= 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} \\ &= \frac{1}{2} \end{aligned}$$

Substituting back into Bayes' theorem, the desired probability is then

$$\begin{aligned} P(\text{rejected}|A) &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{1/2} \\ &= \frac{1}{3} \end{aligned}$$

As expected, my probability of rejection does not change given the new information. Why is the editor wrong? We can see why from the calculation for $P(A)$. Assuming that the editor tells me that A will be published, that event is twice as likely to arise

from the rejection of B (unconditioned probability $1/3$) than from the rejection of my paper (unconditioned probablity $1/6$). It is true that only two papers remain, but B is twice as likely to be rejected.

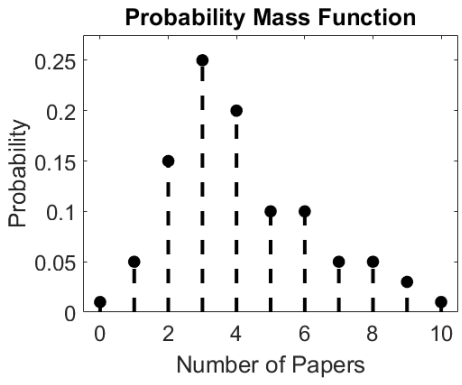
3 Introduction to Probability Distribution Functions

Until this point, we have focused on fundamental rules for defining the probabilities of specific events of interest. We now want to look at *probability distributions*, which define probabilities for every possible outcome of a statistical event. For example, I might want to know the probability distribution of the number of papers published by graduate students during their PhDs or the probability distribution of measurement errors in a given instrument. Below, we will give specific examples of probability distribution functions to illustrate their key properties.

3.1 Probability Mass Function

Let's start by looking at the number of papers published by graduate students during their PhDs. For simplicity, we'll assume a maximum number of 10. Since these values are integers, we are considering a discrete random variable. A *probability mass function* (PMF) defines a probability distribution for a discrete variable. I have given a sample PMF for this case below, defined numerically in the table (left) and represented graphically in the figure (right).

Number of Papers	Probability
0	0.01
1	0.05
2	0.15
3	0.25
4	0.20
5	0.10
6	0.10
7	0.05
8	0.05
9	0.03
10	0.01



What makes this distribution a PMF? First, we have enumerated all possible outcomes (again, assuming a maximum of 10 papers). Second, we have defined valid probabilities p for each of those outcomes, satisfying $0 \leq p \leq 1$. Third, the sum of all of the probabilities is equal to 1 (which follows from the first point). Thus, a PMF gives the probabilities of every possible outcome of a discrete random vari-

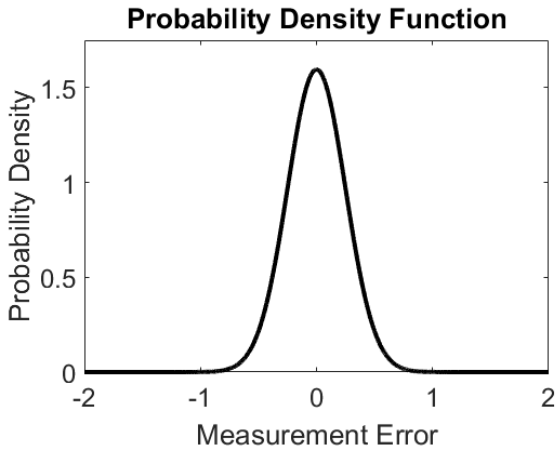
able. A PMF can be defined empirically, as in this case. However, there are also defined distributions like the binomial or Poisson distributions that lend themselves to describing specific types of variables or "stories," as discussed in Tutorial 3c.

3.2 Probability Density Function

A *probability density function* (PDF) is analogous to a PMF, except that a PDF describes a continuous random variable. For example, we might say that measurement errors for a given instrument follow a normal or Gaussian distribution (a prototypical "bell curve") with a mean of $\mu = 0$ and a variance of $\sigma^2 = 0.25$. Mathematically, the probability density that random variable X takes on some value x is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This PDF will be discussed at greater length throughout the course, so don't worry too much about the exact equation now. The PDF is shown graphically below.



What constitutes a PDF? Just like a PMF, the PDF must define probabilities for all possible outcomes (here, all real numbers). You may have noticed that some of the values in the PDF I plotted are actually greater than 1. How can these values be valid probabilities? In the continuous case, the values of the PDF at some given value x don't represent the actual probability that $X = x$; rather, they represent its relative likelihood. In some sense, since X can take on infinite values, the probability that it is exactly equal to a specific value x is infinitesimal, or 0. Instead, we can compute the probability that X takes on values in some range as

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

Thus, the PDF must always be nonnegative, but the values can be greater than 1. Instead, the PDF is constrained by the requirement that the probabilities of all outcomes must total 1, or

$$\int_{-\infty}^{\infty} f(x)\ dx = 1$$

3.3 Cumulative Distribution Function

PMFs and PDFs are ways to specify probability distributions, but that information can also be conveyed in other ways. One example is the *cumulative distribution function* (CDF), defined as

$$F(x) = P(X \leq x)$$

Thus, it defines the probability that X will take any value up to and including x. For a continuous variable, the CDF is the integral of the PDF, or

$$F(x) = \int_{-\infty}^{x} f(x)\ dx$$

Therefore, the CDF approaches 0 as $x \to -\infty$ and 1 as $x \to \infty$.

3.4 Empirical Cumulative Distribution Function

The *empirical cumulative distribution function* (ECDF) is defined exactly the same way as a CDF, except it is applied to empirical data; we might plot the ECDF when trying to measure a continuous variable by taking discrete observations. This way of displaying data shows every measurement without losing any information, unlike plots like a histogram or a box-and-whisker plot.

3.5 Marginal Distribution

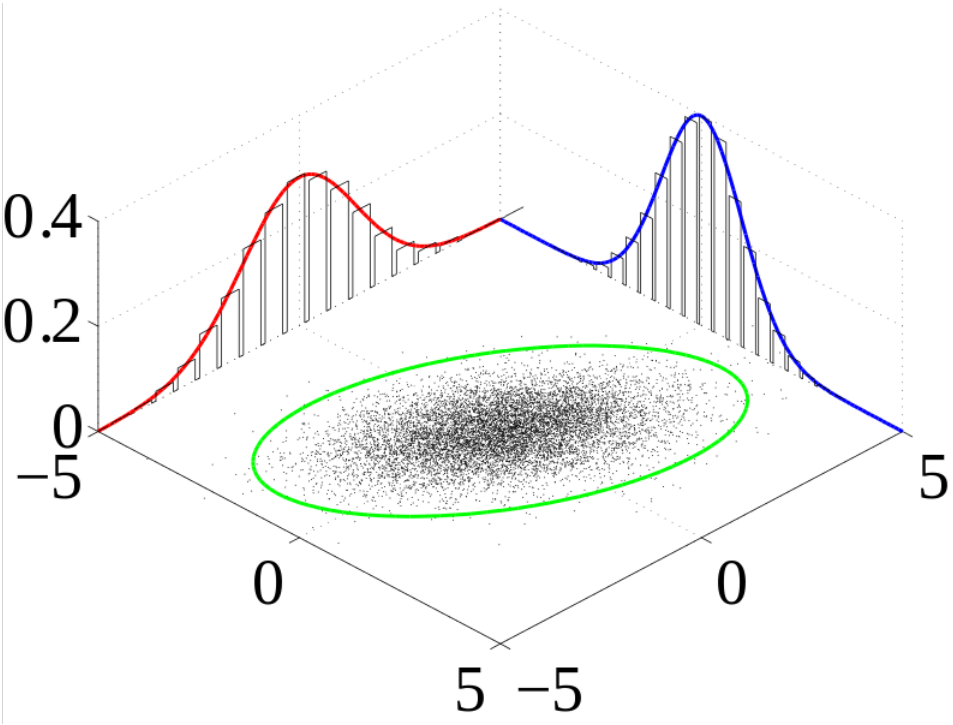
Thus far, we have focused on distributions of a single variable. However, we can also have joint distributions over n variables, with PDF $f(x_1, x_2, \ldots, x_n)$. Often, we want to analyze a particular subset of interest, which we can do by marginalization. If we want the distribution for the first k variables (with $k < n$), we can compute this *marginal distribution* as

$$f(x_1, x_2, \ldots, x_k) = \sum_{x_{k+1}} \cdots \sum_{x_n} f(x_1, x_2, \ldots, x_n)$$

in the discrete case or

$$f(x_1, x_2, \ldots, x_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\ dx_{k+1} \ldots dx_n$$

We sum or integrate over a set of variables (the ones that are marginalized out) in order to fully account for their contributions to the distribution of the remaining variables. As an example, consider the bivariate distribution shown below. Samples from the joint distribution $f(x_1, x_2)$ are shown in green. The marginal distributions $f(x_1)$ and $f(x_2)$ are illustrated in red and blue, representing the distributions of each variable alone.



https://commons.wikimedia.org/wiki/File:Multivariate_normal_sample.svg

3.6 Conditional Distribution

In addition to these distribution functions, we can also have *conditional distributions*, denoted $f(x|y)$. The interpretation is analogous to that of conditional probability, and the relationships between joint and conditional probabilities also hold true for joint and conditional distributions. Consequently, Bayes' theorem holds for distributions, or

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

$$= \frac{f(y|x)f(x)}{f(y)}$$

We can rename x as θ, denoting the model parameter(s) to be estimated, and let y represent the observed or measured data. We then have our usual form for Bayes' theorem, which represents the foundation for the Bayesian inference we will do later in the course.

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

$$= \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

4 Calculation of Summary Statistics

While ECDFs or other probability distribution functions are very informative, they are often not the easiest to interpret. In addition, you may want to capture the information in your data with just a few key values, which you can compare readily across different conditions. To do so, we can consider different *summary statistics* for a set of data. In the real world, we will often see that our data are unimodal, or centered around a single peak. Therefore, we can frequently get a sense for the general shape of a distribution by quantifying the *center*, or some measure of the single peak, and *spread*, or some measure of the distance of the data from the peak. This section provides a brief overview of commonly used summary statistics. See Lecture 4 for a more formal derivation of how these parameters can be estimated from empirical distributions using the plug-in principle.

4.1 Measures of Center

The most common measures of center are the mean and median; the mode can also be considered a measure of center. Specifically, the *mean* of a distribution refers to the expected value or weighted average of that random variable. We will show an example below. The *median* is, intuitively speaking, the "middle" value of a distribution or data set; it is the value at which 50% of the values fall below (or lie above) it. Because the median is defined primarily by the rank of values and not their actual numerical values, this measure is more robust to outliers, or extreme values, than is the mean. Finally, the *mode* is the value that occurs most frequently in a data set. For a continuous distribution, the mode is considered the value at which the probability distribution has a local maximum. The mode works for many unimodal distributions, but it is highly sensitive to skewed distributions. In general, the mean and median are the most common summary statistics for measuring the typical or most representative value of a distribution or data set.

4.1.1 Mean

Because the mean is so commonly used, we will define it formally and give an example of calculating the mean of a distribution. For a discrete random variable X that takes on n values $x_1, x_2, \ldots, x_n$ with PMF $P(x)$, the mean or expected value μ, also denoted $\langle X \rangle$, is

$$\mu = \sum_{i=1}^{n} x_i P(x_i)$$

Thus, we sum over all possible values of the discrete variable and weight each value by its probability. If we are taking the mean of a data set rather than a distribution, we give each observation equal probability or equal weight, so this formula reduces to the arithmetic mean or average, or

$$\begin{aligned}\mu &= \sum_{i=1}^{n} x_i \cdot \frac{1}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i\end{aligned}$$

For a continuous variable with PDF $f(x)$, we convert the summation to integration, or

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

You will nearly always use `numpy` functions or known properties of a distribution rather than performing a calculation by hand. However, just to give a concrete demonstration of using these formulas, we'll give a small example of calculating the mean in practice.

Example: Suppose that I enjoy going to Yogurtland. Each time, I buy \$5 of yogurt; however, every 10th yogurt is free. If X denotes the amount of money I spend at Yogurtland in one visit, I want to know the mean amount μ that I spend, or $\langle X \rangle$. We'll start by defining the PMF. X takes on two values, 0 and 5, with probabilities 0.1 and 0.9. Applying the formula, we find

$$\begin{aligned}\mu &= \sum_{x} x P(x) \\ &= 0 \cdot 0.1 + 5 \cdot 0.9 \\ &= 4.5\end{aligned}$$

Thus, I spend an average of \$4.50 per visit.

4.2 Measures of Spread

The most common measure of spread is the *variance*, which measures how far the values are from the mean. A metric of spread often computed in conjunction with the median is the *interquartile range*, or the differences between the values representing the 25th and 75th percentiles of the data.

4.2.1 Variance

The variance σ^2 of a distribution or data set is defined as

$$\sigma^2 = \langle (X - \langle X \rangle)^2 \rangle$$

In words, the variance is the expected value of the squared distance from the mean. Since expectation, or $\langle X \rangle$, is a linear operator, the variance can also be simplified as

$$\begin{aligned}
\sigma^2 &= \langle (X - \langle X \rangle)^2 \rangle && \text{definition} \\
&= \langle X^2 - 2X\langle X \rangle + \langle X \rangle^2 \rangle && \text{expansion} \\
&= \langle X^2 \rangle - \langle 2X\langle X \rangle \rangle + \langle \langle X \rangle^2 \rangle && \text{linearity} \\
&= \langle X^2 \rangle - 2\langle X \rangle \langle X \rangle + \langle X \rangle^2 && \text{expectation of a constant} \\
&= \langle X^2 \rangle - \langle X \rangle^2 && \text{simplification}
\end{aligned}$$

Example: With this definition, we can now calculate the variance for any arbitrary distribution. Let's calculate the variance of my Yogurtland spending. We first calculate the expected value $\langle X^2 \rangle$ as

$$\begin{aligned}
\langle X^2 \rangle &= \sum_x x^2 P(x) \\
&= 0^2 \cdot 0.1 + 5^2 \cdot 0.9 \\
&= 22.5
\end{aligned}$$

We already computed $\langle X \rangle$ above, so we find that the variance is

$$\begin{aligned}
\sigma^2 &= \langle X^2 \rangle - \langle X \rangle^2 \\
&= 22.5 - 4.5^2 \\
&= 2.25
\end{aligned}$$

The variance of my Yogurtland spending is \$2.25.

5 Conclusion

In this recitation, we have defined the meanings of probabilities and probability distribution functions and shown fundamental rules for calculating and analyzing them. These principles underlie the methods of statistical inference that will be taught in this course, and having a strong foundation in these ideas will help in understanding the "how" and "why" of our approach. In particular, we have seen how Bayes' theorem can be used to quantify how probabilities change as we receive new information, a concept and tool that is applicable to any question or field of interest. Over the next few months, we hope you will get a deeper understanding and appreciation of the power and generality of statistical inference. Happy data analysis!