Discrete Probability

Probability is the key to understanding uncertainty, and thus is essential to many human endeavors, especially the sciences. In this report we will describe the basic qualities we can use to understand discrete events.

Jacob Denson

Chapter 1

Foundations

What is probability? This is a complicated question and cannot be answered without further knowledge of the probabilistic systems we will create. The structures we create will seem very abstract, but this cannot be helped. We will flesh out our intuitive view of these structures as the course goes on.

1.1 Basic Definitions

A sample space is a tuple $(\mathcal{X}, \mathbf{P})$, where $\mathcal{X}$ is the set of outcomes to some experience, and $\mathbf{P}$ is a function from the subsets of $\mathcal{X}$, called events, to the real numbers, called the probability distribution or measure, which satisfies the following properties:

- For any event A, $\mathbf{P}(A) \geq 0$.
- If (A_i) is an infinite sequence of disjoint events, $\mathbf{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)$.
- $\mathbf{P}(\mathcal{X}) = 1$

If you are of the frequentist school, you believe that the sample space is the measurement to some repeatable experiment, and that an event is an observable phenomena. Given an event A and natural number n, define $f(A, n)$ to be the number of times A was seen in n trials of the experiment. Define $p(A, n) = f(A, n)/n$, the relative frequency that the event A was seen. A frequentist would interpret $\mathbf{P}(A)$ to be the relative frequency of the event A after a relatively large number of trials was taken. That is, $\mathbf{P}(A) = \lim_{n\to\infty} p(A, n)$. Then it obviously follows that $f(\mathcal{X}, n) = n$, as $\mathcal{X}$ is the event of some measurement occuring, and thus $\lim_{n\to\infty} p(X, n) = \lim_{n\to\infty} 1 = 1$. If (A_i) is a disjoint sequence of events, it is also intuitive that $f(\cup_{i=1}^{\infty} A_i, n)$ is $\sum_{i=1}^{\infty} f(A_i, n)$, and thus $\mathbf{P}(\cup_{i=1}^{\infty} A_i)$ is equal to $\sum_{i=1}^{\infty} \mathbf{P}(A_i)$. Frequentists assume that the normalized frequency p will converge for all inputs – every experiment has some average occurence. $A \cap B$ is seen as the event where A and B occurs, and $A \cup B$ is the event where A or B occurs.

Note that this is not a mathematic definition of probability, but instead an instead an interpretation of the axiomatic foundations. Another, more complicated interpretation is the bayesian, which will come later. It is also important to see that just because the probability of some event is 0 does not imply that A is impossible, just that is almost certainly will not happen (Just because a sequence converges to 0 does not imply the sequence is 0 everywhere). Similarily, just because the probability of some event is 1 does not imply that the measurement will always occur in an experiment. The only certain event is $\mathcal{X}$, and the only impossible event is $\emptyset$.

Using the basic axioms of probability theory, some easy properties of probabilities occur:

- For any event A, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$:

  Proof. A and A^c are disjoint, and the union of the two is X □

- $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$

  Proof. $\mathcal{X} = A \cup A^c$ □

- $\mathbf{P}(\emptyset) = 0$

  Proof. $\emptyset = \mathcal{X}^c$ □

- If A is a subset of B, $\mathbf{P}(A) \leq \mathbf{P}(B)$

  Proof. The property follows as $\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B - A)$, and $\mathbf{P}(B - A) \geq 0$ □

- For any events A and B, $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$:

  Proof. $A \cup B$ is the union of three disjoint events $A \cap B^c$, $B \cap A^c$, and $A \cap B$. The following calculation results in the equation:

$$\begin{aligned}
\mathbf{P}(A \cup B) &= \mathbf{P}(A \cap B^c) + \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B) \\
&= \mathbf{P}(A \cap B^c) + \mathbf{P}(A \cap B) + \mathbf{P}(A^c \cap B) + \mathbf{P}(A \cap B) - \mathbf{P}(A \cap B) \\
&= \mathbf{P}((A \cap B^c) \cup (A \cap B)) + \mathbf{P}((A^c \cap B) \cup (A \cap B)) - \mathbf{P}(A \cap B) \\
&= \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)
\end{aligned}$$

  □

Less obvious is the following theorem, called 'the continuity of probabilities'. Let (A_i) be a sequence of events such that $A_i \subseteq A_j$ for all $j \geq i$. Then $\mathbf{P}(\cup_{i=1}^{\infty} A_i) = \lim_{n \to \infty} A_n$:

Proof. Give the sequence (A_i), define a new sequence (B_i) recursively by $B_1 = A_1$, and $B_n = A_n - \cup_{i=1}^{n-1} B_i$. These are disjoint and there is union is the same as the union of A_i, so $\mathbf{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbf{P}(B_i) = \lim_{n\to\infty} A_n$. □

It is easy to dualize the continuity of probabilities. If the set sequence is decreasing, the same property holds for intersections.

1.2 Conditionality

Given some information about the sample space, we should be able to infer that other probabilities change. This motivates the conditional probability. Given some event A such that $\mathbf{P}(A) > 0$, we define the probability that some event B happens given that A happens, denoted $\mathbf{P}(B|A)$, to be $\mathbf{P}(A \cap B)/\mathbf{P}(A)$. $\mathbf{P}(B)$ is the prior probability. Note the function $\mathbf{P}(\cdot|A)$ defines a new probability distribution on the sample space – it is the probability of the world after some event occurs.

Under frequentist interpretation, $\mathbf{P}(B|A)$ is the relative frequency of event A being measured when B is measured, to B being measured, $\lim_{n\to\infty} f(A \cap B, n)/f(B, n)$. But this is equivalent to $\lim_{n\to\infty}(f(A \cap B, n)/n)/(f(B, n)/n)$, which is $\lim_{n\to\infty} \mathbf{P}(A \cap B)/\mathbf{P}(A)$. The frequentist interpretation is equivalent to the mathematical definition.

We say two events A and B are independent if $\mathbf{P}(A|B) = \mathbf{P}(A)$. Intuitively, knowledge that B happened tells us nothing about the relative occurence of A. As $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$, and event is independent if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

As an example, consider two coin flips on a fair coin. We should not be able to infer information about one coin flip from the result of the other. Let the sample space X be the outcomes of an experiment. Then $\mathcal{X} = \{H, T\}^2$, where the first coordinate of the cartesian product is the result of the first flip of the coin, and the second coordinate the second result. Experiments affirm that each singular event in the sample space is equiprobable, according to the frequentist interpretation. Then the event that our first coin flip got a heads is $A = \{(H, H), (H, T)\}$. The event that the second flip was a tails is $B = \{(H, T), (T, T)\}$. The event where both happen is $\{(H, T)\}$. Then $\mathbf{P}(B|A) = \mathbf{P}(\{(H, T)\})/\mathbf{P}(\{(H, T), (H, H)\{) = (1/4)/(2/4) = 1/2$, and $\mathbf{P}(B) = 1/2$, so A and B are independent – Our intuitions are affirmed.

The sample space of the previous example is an instance of a more general idea. Given two sample spaces $\mathcal{X}$ and Ω, with probability measures $\mathbf{P}_{\mathcal{X}}$ and $\mathbf{P}_{\Omega}$ we define the product space $(\mathcal{X} \times \Omega)$ with probability measure extended from $\mathbf{P}_{\mathcal{X}\times\Omega}(A \times B) = \mathbf{P}_{\mathcal{X}}(A)\mathbf{P}_{\Omega}(B)$. This of course means that the two sets $A \times \Omega$ and $\mathcal{X} \times B$ are independent for all events B.

One theorem which helps us understand sample space is the law of total probability. Let $A_1, A_2, \ldots$ be disjoint events that partition the sample space, where the probability of each is greater than 0. Then for any event B, $\mathbf{P}(B) = \sum_{i=1}^{\infty} \mathbf{P}(B|A_i)\mathbf{P}(A_i)$. The proof is self explanatory.

From the law of total probability, we obtain Bayes theorem. Consider B and A_i from the previous theorem. Then $\mathbf{P}(A_i|B) = \mathbf{P}(B|A_i)\,\mathbf{P}(A_i)/\sum_{j=1}^{\infty}\mathbf{P}(B|A_j)\,\mathbf{P}(A_j)$.

Proof. We obtain the equation as $\mathbf{P}(A_i|B) = \mathbf{P}(A_i \cap B)/\mathbf{P}(B)$. To find the denominator, $\mathbf{P}(B) = \sum_{i=1}^{\infty}\mathbf{P}(B|A_i)\,\mathbf{P}(A_i)$, and the numerator by $\mathbf{P}(A_i \cap B) = \mathbf{P}(B|A_i)\,\mathbf{P}(A_i)$. □

We can generalize independence to an arbitrary collection of sets $\{A_i\}_{i\in I}$ where I is an index set. Then this collection is independent if $\mathbf{P}(\cap_{i=a_1}^{a_n} A_i) = \prod_{i=1}^{n}(A_i)$ for any subset $A_{a_1}, A_{a_2}, \ldots, A_{a_n}$.

It is common for a beginner in probability theory to think that if two events A and B are disjoint, then the two events are independent. This coudn't be further from the truth. Think about it this way. If A has happened, B cannot happened, so we know the posterior probability of B to be 0 after A. Absolute information is obtained from knowledge of A.

There are some useful inequalities for working with probability spaces. The first is boole's inequality, that $\mathbf{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty}\mathbf{P}(A_i)$. Another is bonferroni's inequality, that $\mathbf{P}(\cup_{i=1}^{n} A_i) \geq \sum_{i=1}^{n}\mathbf{P}(A_n) - \sum_{r<i}\mathbf{P}(A_r \cap A_k)$.

Chapter 2

Random Variables

Common to the sciences is the notion of an experimental value, a measurement of a number that summarizes that a specific event has happened. We define this mathematically with the concept of a random variable. A random variable is a mapping from the sample space to the real numbers. Intuitively, think of it as, when a measurement is obtained from the sample space, it corresponds to a real number that the random variable is mapped to. The importance of the random variable is that it infers a sample space ($\mathbf{R}, \mathbf{P}(X \in \cdot)$, defined by $\mathbf{P}(X \in \cdot) = \mathbf{P}(X^{-1}(\cdot))$. It is useful to define, for some logical statement R, $P(R(X)) = \mathbf{P}(\{x \in \mathbf{R} | R(x)\})$. Some example uses are $\mathbf{P}(X = 2)$, $\mathbf{P}(X < 2)$.

Given a random variable X, we define the cumulative distribution function $F_X(x) = \mathbf{P}(X \leq x)$. This function has useful properties, some so trivial that the proof is not shown:

- $\mathbf{P}(a < X \leq b) = F_X(b) - F_X(a)$
- $\mathbf{P}(X > a) = 1 - F_X(a)$
- $\lim_{x \to -\infty} F_X(x) = 0$

  Proof. Consider the sequence of sets $\{P_i\}$ in the sample space $\mathcal{X}$ where $P_i = \{x \in \mathcal{X} | X(x) \leq -i\}$. $\{P_i\}$ is decreasing, so $\lim_{i \to \infty} P_i = \mathbf{P}(\cap_{i=1}^{\infty} P_i)$, which is $\mathbf{P}(\emptyset) = 0$. As F_X is a decreasing function, this implies the function converges to the same value. □

- $\lim_{x \to \infty} F_X(x) = 1$

  Proof. Use the same method as the previous proof □

To make life easy for the rest of the book, we now assume all random variables are discrete, that is, the range of real numbers they take on is countable. Though it is possible to continue the theory of probability on arbitrary random variables, we require the complex topic of measure theory to be able to manage

the strange ways that uncountable ranges can have on the variables. We leave this for a later course.

We also define the inverse of F_X, the quantile function $F_X^{-1}(x)$ to be the infinum of the set $\{q|F_X(q) \leq x\}$, defined on $[0,1]$. We call $F_X^{-1}(1/4)$ the first quartile, $F_X^{-1}(1/2)$ the second quartile, and so on.

Given a random variable X, define the probability mass function $f_X(x) = \mathbf{P}(X = x)$. We have that $F_X(x) = \sum_{a \leq x} f_X(a)$ (see where we have a problem with uncountably many values for the random variable). Furthermore, $f_X(x) = F_X(x) - \lim_{y \to x^-} F_X(y)$. These facts lead to the intricate connection between the mass function and the distribution function, completely defining a random variable in the following way. Let X and Y be two random variables (which may or may not be in the same sample space). Then $F_Y = F_X$ if and only if $f_X = f_Y$, and either holds, we say X and Y share the same distribution, or $X \sim Y$:

Proof. We have that $f_X(x) = F_X(x) - \lim_{y \to x^-} F_X(y)$, so if $F_x = F_y$, $F_X(x) = F_Y(x)$ for all x. Likewise, if $f_X = f_Y$, as $F_X(x) = \sum_{a \leq x} f_X(a)$, the two cumulative distribution functions are equal. □

2.1 Distributions

In order to get some practice using distributions, we list some very common discrete distributions everyone should know, along with some explanations and symbols used for them.

The Point Mass distribution δ_a is such that when $X \sim \delta_a$, $\mathbf{P}(X = a) = 1$, and $\mathbf{P}(X \neq a) = 0$.

The discrete or uniform distribution $\text{Uni}(k)$ is defined on integers from 1 to k, by $\mathbf{P}(X = x) = 1/k$.

The Bernoulli distribution $\text{Ber}(p)$ represents flipping a coin with a probability p of getting a heads, and $1 - p$ of getting a tails, and assigning a value of one for a heads and zero for tails. For a random variable X, we say $X \sim \text{Ber}(p)$ if $\mathbf{P}(X = 1) = p$ and $\mathbf{P}(X = 0) = 1 - p$.

The Binomial distribution $\text{Bin}(p, n)$ represents flipping a coin like in the Bernoulli distribution n times, and counting how many times a head was gotten. Note the a $\text{Ber}(p)$ distribution is equivalent to a $\text{Bin}(p, 1)$ distribution. The mass function f_X of a random variable X such that $X \sim \text{Bin}(p, n)$ is defined by the map $\mathbf{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ for an integer x in the range 1 to n.

The Geometric distribution $\text{Geo}(p)$ represents flipping a coin until a heads turned up, and counting how many flips it took. The mass function is defined on $\mathbf{N}^+$ by $f_X(x) = p(1-p)^{x-1}$.

The Poisson distribution $\text{Poisson}(\lambda)$ represents the number of independant events that occur in a fixed period of time given the average number of events that occur is λ. The mass function is defined on $\mathbf{N}$ by $f_X(x) = e^{-\lambda} \lambda^x / x!$.

Note that in none of these distributions have we defined a sample space for the random variables used. This is because in most circumstances the sample

space does not matter once the distribution of the random variables – this is one of the elegances of the definition.

Given two random variables X and Y, we may define a new sample space over $\mathbf{R}^2$ by $\mathbf{P}(x,y) = \mathbf{P}(X = x, Y = y)$. We define the joint mass distribution $f_{X,Y}(x,y) = \mathbf{P}(x,y)$. We call this a bivariate distribution.

This allows us to talk about mappings on $\mathbf{R}^2$ random variables in terms of variables on $\mathbf{R}$, but how do we get to $\mathbf{R}$ from a distribution over $\mathbf{R}^2$. The trick is to obtain (X,Y) by the marginal mass function $f_X(x) = \mathbf{P}(\{x\} \times \mathbf{R})$, and $f_Y(y)\,\mathbf{P}(\mathbf{R} \times \{y\})$.

Given a join distribution mass function $f_{X,Y}$, we say X and Y are independant, and write $X \perp\!\!\!\perp Y$, if $f_{X,Y} = f_X f_Y$. An example of independance is the following. Let $X \sim \text{Bin}(n,p)$. Then $X = X_1 + X_2 + \cdots + X_n$, where $X_i \perp X_j$ for $i \neq j$, and $X_i \sim \text{Ber}(p)$. Think of each X_i as a single coin flip.

We can generalize the concept of a bivariate distribution to an arbitrary number of random variables, a multivariate distribution. A random vector is a vector of random variables $X = (X_1, X_2, \ldots, X_n)$. We define the probability by $\mathbf{P}(x_1, x_2, \ldots, x_n) = \mathbf{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. We say these variables are independant if $f_X = f_{X_1} f_{X_2} \cdots f_{X_n}$.

The main discrete random vector distribution is the multinomial distribution $\text{mult}(n,p)$, where $p = (p_1, p_2, \ldots, p_n)$ for $p_i \in \mathbf{N}$. See this as drawing n balls from an urn where there are p_i balls of colour i. Then for a vector $x = (x_1, x_2, \ldots, x_n)$, such that $\sum_{i=1}^n x_i = n$ define the mass function $f_X(x) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}$. Here $\binom{n}{x_1, x_2, \ldots, x_n} = n!/(x_1! x_2! \ldots x_n!)$.

2.2 Expectation

Given a set of experimental results, we may take the average of the result hoping to obtain a new result closest to the average of the result of taking every value from the experiment. This is the expectation of a random variable $\mathbf{E}(X) = \sum_x x\,\mathbf{P}(X = x)$, summed over the range of a random variable. Think of $\mathbf{E}\,E$ as the summary of the distribution. We calculate some expectations below:

- If $X \sim \text{Ber}(p)$, then $\mathbf{E}(X) = (0)(1-p) + (1)(p) = p$
- If $X \sim \text{Bin}(n,p)$, then $\mathbf{E}(X) = \sum_{k=0}^n k p^k (1-p)^{n-k} n!/((n-k)!k!)$. This is a hard sum to calculate, so we use a theorem we will prove later by noting that $X = X_1 + \cdots + X_n$ where $X_i \sim \text{Ber}(p)$, and thus $\mathbf{E}(X) = \sum_{k=0}^n \mathbf{E}(X_i) = np$.
- If $X \sim \delta_a$, $\mathbf{E}(X) = a$.
- If $X \sim \text{Uni}(k)$, $\mathbf{E}(X) = \sum_{n=1}^k n/k = k(k+1)/2k = (k+1)/2$.
- If $X \sim \text{Geo}(p)$, $\mathbf{E}(X) = \sum_{x=1}^\infty xp(1-p)^{x-1} = p/1 - p\sum_{x=1}^\infty x(1-p)^x$, which is equal to $p/(1-p)\sum_{i=1}^\infty \sum_{j=i}^\infty (1-p)^x$, which is $p/(1-p)\sum_{i=1}^\infty 1/p\sum_{i=1}^n (1-p)^i$ which comes out to be p.

- If $X \sim \text{Poisson}(\lambda)$, $\mathbf{E}(X) = \sum_{i=1}^{\infty} x e^{-\lambda} \lambda^x / x! = e^{-\lambda} \sum \lambda^x / (x-1)! = \lambda e^{-\lambda} \sum \lambda^{x-1} / (x-1)!$, which turns out to be $\lambda e^{-\lambda} e^{\lambda} = \lambda$.