Probability Distributions

Throughout the past chapters, we've actually already encountered many of the topics in this section. In order to define things like expectation and variance, we introduced random variables denoted X or Y as mappings from the sample space to the real numbers. All of the distributions we've so far looked at have been what are called *discrete* distributions. We will soon look at the distinction between discrete and continuous distributions. Additionally we will introduce perhaps the most influential theorem in statistics, the *Central Limit Theorem*, and give some applications.

Random Variables

In Section 2.2 (Expectation), we wanted to find the expectation of a coin flip. Since the expectation is defined as a weighted sum of outcomes, we needed to turn the outcomes into numbers before taking the weighted average. We provided the mapping

$$\mathsf{T} \mapsto 0$$
$$\mathsf{H} \mapsto 1$$

Here was our first encounter of a random variable.

Definition 0.0.11. *A function X that maps outcomes in our sample space to real numbers, written $X : \Omega \to$, is called a* ***random variable****.*

In the above example, our sample space was

$$\Omega = \{\mathsf{H}, \mathsf{T}\}$$

and our random variable $X : \Omega \to$, i.e. our function from the sample space Ω to the real numbers , was defined by

$$X(\mathsf{T}) = 0$$
$$X(\mathsf{H}) = 1$$

Now would be a great time to go onto the website and play with the "Random Variable" visualization. The sample space is represented

by a hexagonal grid. Highlight some hexagons and specify the value your random variable X assigns to those hexagons. Start sampling on the grid to see the empirical frequencies on the left.

Independence of Random Variables

In previous sections we've mentioned independence of random variables, but we've always swept it under the rug during proofs since we hadn't yet formally defined the concept of a random variable. Now that we've done so, we can finally define a second form of independence (different from independence of *events*).

Definition 0.0.12. *Suppose X and Y are two random variables defined on some sample space Ω. We say X and Y are* ***independent random variables*** *if*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any two subsets A and B of .

Let's go back and prove Exercise 2.9 (c), i.e. that if X and Y are independent random variables, then

$$E[XY] = E[X]E[Y]$$

Proof. Define the random variable $Z(\omega) = X(\omega)Y(\omega)$. By the definition of expectation, the left hand side can be written

$$\begin{aligned}
E[XY] &= \sum_{z \in Z(\Omega)} z \cdot P(Z = z) \\
&= \sum_{x \in X(\Omega), y \in Y(\Omega)} xyP(X = x, Y = y) \\
&= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xyP(X \in \{x\}, Y \in \{y\}) \\
&= \sum_{x \in X(\Omega)} \sum_{y \in Y(\Omega)} xyP(X \in \{x\})P(Y \in \{y\}) \\
&= \sum_{x \in X(\Omega)} xP(X \in \{x\}) \sum_{y \in Y(\Omega)} yP(Y \in \{y\}) \\
&= E[X]E[Y]
\end{aligned}$$

This completes the proof. □

Discrete vs. Continuous

Thus far we have only studied discrete random variables, i.e. random variables that take on only up to *countably* many values. The word "countably" refers to a property of a set. We say a set is *countable* if we can describe a method to list out all the elements in the set such that for any particular element in the set, if we wait long enough in our listing process, we will eventually get to that element. In contrast, a set is called *uncountable* if we cannot provide such a method.

Countable vs. Uncountable

Let's first look at some examples.

Example 0.0.2. *The set of all natural numbers*

$$\mathbb{N} \doteq \{1, 2, 3, \dots\}$$

is countable. Our method of enumeration could simply be to start at 1 and add 1 every iteration. Then for any fixed element $n \in \mathbb{N}$, this process would eventually reach and list out n.

Example 0.0.3. *The integers,*

$$\mathbb{Z} \doteq \{0, 1, -1, 2, -2, 3, -3, \dots\}$$

is countable. Our method of enumeration as displayed above is to start with 0 for the first element, add 1 to get the next element, multiply by -1 to get the third element, and so on. Any integer $k \in \mathbb{Z}$, if we continue this process long enough, will be reached.

Example 0.0.4. *The set of real numbers in the interval $[0, 1]$ is uncountable. To see this, suppose for the sake of contradiction that this set were countable. Then there would exist some enumeration of the numbers in decimal form. It might look like*

$$\begin{array}{l} 0.1354295\dots \\ 0.4294726\dots \\ 0.3916831\dots \\ 0.9873435\dots \\ 0.2918136\dots \\ 0.3716182\dots \\ \vdots \end{array}$$

Consider the element along the diagonal of such an enumeration. In this case the number is

$$a \doteq 0.121318\dots$$

Now consider the number obtained by adding 1 to each of the decimal places, i.e.

$$a' \doteq 0.232429\ldots$$

This number is still contained in the interval $[0,1]$*, but does not show up in the enumeration. To see this, observe that* a' *is not equal to the first element, since it differs in the first decimal place by 1. Similarly, it is not equal to the second element, as* a' *differs from this number by 1 in the second decimal place. Continuing this reasoning, we conclude that* a' *differs from the* n^{th} *element in this enumeration in the* n^{th} *decimal place by 1. It follows that if we continue listing out numbers this way, we will never reach the number* a'*. This is a contradiction since we initially assumed that our enumeration would eventually get to every number in* $[0,1]$*. Hence the set of numbers in* $[0,1]$ *is uncountable.*

If you're left feeling confused after these examples, the important take away is that an uncountable set is *much* bigger than a countable set. Although both are infinite sets of elements, uncountable infinity refers to a "bigger" notion of infinity, one which has no gaps and can be visualized as a continuum.

Discrete Distributions

Definition 0.0.13. *A random variable X is called* ***discrete*** *if X can only take on finitely many or countably many values.*

For example, our coin flip example yielded a random variable X which could only take values in the set $\{0,1\}$. Hence, X was a discrete random variable. However, discrete random variables can still take on infinitely many values, as we see below.

Example 0.0.5 (Poisson Distribution)**.** *A useful distribution for modeling many real world problems is the Poisson Distribution. Suppose* $\lambda > 0$ *is a positive real number. Let X be distributed according to a Poisson distribution with parameter* λ*, i.e.*

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

where $k \in \mathbb{N}$*. The shorthand for stating such a distribution is* $X \sim Poi(\lambda)$*. Since k can be any number in* $\mathbb{N}$*, our random variable X has a positive probability on infinitely many numbers. However, since* $\mathbb{N}$ *is countable, X is still considered a discrete random variable.*

On the website there is an option to select the "Poisson" distribution in order to visualize its probability mass function. Changing the value of λ *changes the probability mass function, since* λ *shows up in the probability*

expression above. Drag the value of λ from 0.01 up to 10 to see how varying λ changes the probabilities.

Example 0.0.6 (Binomial Distribution)**.** *Another useful distribution is called the Binomial Distribution. Consider n coin flips, i.e. n random variables $X_1, \ldots, X_n$ each of the form*

$$X_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Now consider the random variable defined by summing all of these coin flips, i.e.

$$S \doteq \sum_{i=1}^{n} X_i$$

We might then ask, "What is the probability distribution of S?" Based on the definition of S, it can take on values from 0 to n, however it can only take on the value 0 if all the coins end up tails. Similarly, it can only take on the value n if all the coins end up heads. But to take on the value 1, we only need one of the coins to end up heads and the rest to end up tails. This can be achieved in many ways. In fact, there are $\binom{n}{1}$ ways to pick which coin gets to be heads up. Similarly, for $S = 2$, there are $\binom{n}{2}$ ways to pick which two coins get to be heads up. It follows that for $S = k$, there are $\binom{n}{k}$ ways to pick which k coins get to be heads up. This leads to the following form,

$$P(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The p^k comes from the k coins having to end up heads, and the $(1-p)^{n-k}$ comes from the remaining $n-k$ coins having to end up tails. Here it is clear that k ranges from 0 to n, since the smallest value is achieved when no coins land heads up, and the largest number is achieved when all coins land heads up. Any value between 0 and n can be achieved by picking a subset of the n coins to be heads up.

Selecting the "Binomial" distribution on the website will allow you to visualize the probability mass function of S. Play around with n and p to see how this affects the probability distribution.

Continuous Distributions

Definition 0.0.14. *We say that X is a **continuous** random variable if X can take on uncountably many values.*

If X is a continuous random variable, then the probability that X takes on any particular value is 0.

Example 0.0.7. *An example of a continuous random variable is a Uniform[0,1] random variable. If $X \sim$ Uniform[0,1], then X can take on any value in the interval [0,1], where each value is equally likely. The probability that X takes on any particular value in $[0,1]$, say $\frac{1}{2}$ for example, is 0. However, we can still take probabilities of subsets in a way that is intuitive. The probability that x falls in some interval (a,b) where $0 \leq a < b \leq 1$ is written*

$$P(X \in (a,b)) = b - a$$

The probability of this event is simply the length of the interval (a,b).

A continuous random variable is distributed according to a *probability density function*, usually denoted f, defined on the domain of X. The probability that X lies in some set A is defined as

$$P(X \in A) = \int_A f$$

This is informal notation but the right hand side of the above just means to integrate the density function f over the region A.

Definition 0.0.15. *A **probability density function** f (abbreviated **pdf**) is valid if it satisfies the following two properties.*

1. $f(x) \geq 0$ *for all* $x \in$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Example 0.0.8 (Exponential Distribution)**.** *Let $\lambda > 0$ be a positive real number. Suppose X is a continuous random variable distributed according to the density*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Let's check that f defines a valid probability density function. Since $\lambda > 0$ and e^y is positive for any $y \in$, we have $f(x) \geq 0$ for all $x \in$. Additionally, we have

$$\begin{aligned} \int_0^{\infty} f(x)dx &= \int_0^{\infty} \lambda e^{-\lambda x} \\ &= \left[\lambda \frac{-1}{\lambda} e^{-\lambda x}\right]_0^{\infty} \\ &= 0 - (-1) \\ &= 1 \end{aligned}$$

Since f is nonnegative and integrates to 1, it is a valid pdf.

Example 0.0.9 (Normal Distribution)**.** *We arrive at perhaps the most known and used continuous distributions in all of statistics. The Normal distribution is specified by two parameters, the mean μ and variance σ^2. To say X is a random variable distributed according to a Normal distribution with mean μ and variance σ^2, we would write $X \sim N(\mu, \sigma^2)$. The corresponding pdf is*

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

Some useful properties of normally distributed random variables are given below.

Proposition 0.0.2. *If $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ are independent random variables, then*

(a) The sum is normally distributed, i.e.

$$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

(b) Scaling by a factor $a \in$ results in another normal distribution, i.e. we have

$$aX \sim N(a\mu_x, a^2\sigma_x^2)$$

(c) Adding a constant $a \in$ results in another normal distribution, i.e.

$$X + a \sim N(\mu_x + a, \sigma_x^2)$$

Heuristic. In order to rigorously prove this proposition, we need to use moment generating functions, which aren't covered in these notes.

However, if we believe that $X + Y$, aX, and $X + a$ are all still normally distributed, it follows that the specifying parameters (μ and σ^2) for the random variables in (a), (b), and (c) respectively are

$$\begin{aligned} E(X + Y) &= EX + EY = \mu_x + \mu_y \\ \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) = \sigma_x^2 + \sigma_y^2 \end{aligned}$$

and

$$\begin{aligned} E(aX) &= aEX = a\mu_x \\ \text{Var}(aX) &= a^2\text{Var}(X) = a^2\sigma_x^2 \end{aligned}$$

and

$$\begin{aligned} E(X + a) &= EX + a = \mu_x + a \\ \text{Var}(X + a) &= \text{Var}(X) + \text{Var}(a) = \text{Var}(X) = \sigma_x^2 \end{aligned}$$

□

The Central Limit Theorem

We return to dice rolling for the moment to motivate the next result. Suppose you rolled a die 50 times and recorded the average roll as $\bar{X}_1 = \frac{1}{50}\sum_{k=1}^{50} X_k$. Now you repeat this experiment and record the average roll as $\bar{X}_2$. You continue doing this and obtain a sequence of sample means $\{\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots\}$. If you plotted a histogram of the results, you would begin to notice that the $\bar{X}_i$'s begin to look normally distributed. What are the mean and variance of this approximate normal distribution? They should agree with the mean and variance of $\bar{X}_i$, which we compute below. Note that these calculations don't depend on the index i, since each $\bar{X}_i$ is a sample mean computed from 50 independent fair die rolls. Hence we omit the index i and just denote the sample mean as $\bar{X} = \frac{1}{50}\sum_{k=1}^{50} X_k$.

$$\begin{aligned} E(\bar{X}) &= E\Big(\frac{1}{50}\sum_{k=1}^{50} X_k\Big) \\ &= \frac{1}{50}\sum_{k=1}^{50} E(X_k) \\ &= \frac{1}{50}\sum_{k=1}^{50} 3.5 \\ &= \frac{1}{50}\cdot 50 \cdot 3.5 \\ &= 3.5 \end{aligned}$$

where the second equality follows from linearity of expectations, and the third equality follows from the fact that the expected value of a die roll is 3.5 (See Section 2.2). The variance of $\bar{X}_i$ is

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \mathrm{Var}\Big(\frac{1}{50}\sum_{k=1}^{50} X_k\Big) && \text{(Definition of } \bar{X}_i\text{)} \\ &= \frac{1}{50^2}\mathrm{Var}\Big(\sum_{k=1}^{50} X_k\Big) && (\mathrm{Var}(cY) = c^2\mathrm{Var}(Y)) \\ &= \frac{1}{50^2}\sum_{k=1}^{50} \mathrm{Var}(X_k) && (X_k\text{'s are independent.}) \\ &= \frac{1}{50^2}\cdot 50 \cdot \mathrm{Var}(X_k) && (X_k\text{'s are identically distributed.}) \\ &\approx \frac{1}{50}\cdot 2.92 \\ &\approx 0.0583 \end{aligned}$$

where we computed $\mathrm{Var}(X_k) \approx 2.92$ in Exercise 2.12. So we would begin to observe that the sequence of sample means begins to resemble a normal distribution with mean $\mu = 3.5$ and variance

$\sigma^2 = 0.0582$. This amazing result follows from the Central Limit Theorem, which is stated below.

Theorem 0.0.4 (Central Limit Theorem)**.** *Let $X_1, X_2, X_3, \ldots$ be iid (independent and identically distributed) with mean μ and variance σ^2. Then*

$$\bar{X} \to N\Big(\mu, \frac{\sigma^2}{n}\Big)$$

in distribution as $n \to \infty$.

All this theorem is saying is that as the number of samples n grows large, independent observations of the sample mean $\bar{X}$ look as though they were drawn from a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$. The beauty of this result is that this type of convergence to the normal distribution holds for any underlying distribution of the X_i's. In the previous discussion, we assumed that each X_i was a die roll, so that the underlying distribution was discrete uniform over the set $\Omega = \{1, 2, 3, 4, 5, 6\}$. However, this result is true for any underlying distribution of the X_i's.

A continuous distribution we have not yet discussed is the Beta distribution. It is characterized by two parameters α and β (much like the normal distribution is characterized by the parameters μ and σ^2.) On the Central Limit Theorem page of the website, choose values for α and β and observe that the sample means look as though they are normally distributed. This may take a while but continue pressing the "Submit" button until the histogram begins to fit the normal curve (click the check box next to "Theoretical" to show the plot of the normal curve).

Corollary 0.0.2. *Another way to write the convergence result of the Central Limit Theorem is*

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

Proof. By the CLT, $\bar{X}$ becomes distributed $N(\mu, \frac{\sigma^2}{n})$. By Proposition 4.14 (c), $\bar{X} - \mu$ is then distributed

$$\bar{X} - \mu \sim N\Big(\mu - \mu, \frac{\sigma^2}{n}\Big) = N\Big(0, \frac{\sigma^2}{n}\Big)$$

Combining the above with Proposition 4.14 (a), we have that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is distributed

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N\Big(0, \frac{\sigma^2}{n} \cdot \Big(\frac{1}{\sigma/\sqrt{n}}\Big)^2\Big) = N(0, 1)$$

□