Gov 2000 - 3. Random Variables and Probability Distributions

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

WHERE ARE WE? WHERE ARE WE GOING?

- We've learned about the fundamentals of probability and worked with events.
- Now we are going to focus on random variables and their distributions. These variables are important because they are direct connections to the data we have in spreadsheets. We want to learn about the distribution of these variables.

LIST EXPERIMENTS FOR SENSITIVE QUESTIONS

- A list experiment is one where we ask respondents to tell us how many items on a list they agree with, where the contents of the list are randomized across respondents.
- For instance, perhaps we want to know what proportion of people would be upset by a black family moving in next door to them. What we would do is the following.
- Randomly split the survey into two halves. In the first half, ask respondents how many of the following upset you:

1. the federal government increasing the tax on gasoline;
2. professional athletes getting million-dollar salaries;

3. large corporations polluting the environment.

- The other half received the same prompt, except with an additional item:

1. the federal government increasing the tax on gasoline;
2. professional athletes getting million-dollar salaries;
3. large corporations polluting the environment;
4. a black family moving in next door.

- It turns out that we can use the answers to these questions to figure out what proportion of the population would be upset by a black family moving in next door. But in order to do that, we need to understand random variables.

WHAT ARE RANDOM VARIABLES?

- Very simply, random variables are functions that map outcomes of the experiment to numbers. Sometimes this connection is obvious or trivial because the sample space is already a collection of numbers. Other times, we need to construct random variables.
- A random variable is a function that maps from the sample space of an experiment to the real line or $X : \Omega \to \mathbb{R}$.
- Imagine our experiment was tossing a coin 5 times. The sequence of outcomes of the flips, $\omega = HTHTT$ for example, is not a random variable because it isn't a number. But we could make it into a random variable, X if make it the number of heads in the 5 tosses. Notice how the random variable takes in outcomes from the sample space ($HTHTT$) and converts them into a number, 2.
- Each sample space can have many different random variables defined on it. For the coin toss, we could also define a variable to be the number of tails flips, Y. In this case, these two variables would be related by $X = 5 - Y$.
- We almost always use capital roman letters for the "name" of the random variable such as X. Here that is just shorthand for the number of heads in 5 coin flips. Obviously when we need to do mathematical operations on the variable, its shorthand name X will be easier to use. We will refer to a particular value that X might take with lower case letters, x. So we might write $\mathbb{P}(X = x)$ to be the probability that the number of heads is equal to x.
- Another experiment might be someone turns out to vote. For just one person, the sample space is $\Omega =$ {voted,didn't vote}. But again these outcomes can be results of a random variable because they are not numeric. We could define a

random variable, X, that converts these outcomes into numbers:

$$X = \begin{cases} 1 \text{ if voted} \\ 0 \text{ if didn't vote} \end{cases}$$

- This last variable is called a **Bernoulli** or **binary** random variable.
- Sometimes the sample space is already numeric so creating random variables is more obvious. What if our experiment is how long a government lasts in a parliamentary system? Obviously here the sample space is the set of nonnegative numbers $\Omega = [0, \infty)$. Then our random variable might just be equal to the outcome itself.

WHY RANDOM VARIABLES?

- Why do we need to introduce these functions? Remember that we said that statistics is the mathematical study of data. In order to use the tools of math to tackle our questions of interest, we are going to need to work with numerical outputs.
- Working with the original sample space might be incredibly difficult and very application specific. But once we convert these sample spaces into random variables, we can see that very different problems might lead to random variables with very similar properties.

Why are these variables random?

- It might seem confusing at first that we call these random variables since they deterministically map from the sample space to the real line. Where does the randomness come from?
- One way to view random variables is that they are numerical summaries of the experiments that occurred. Our uncertainty over which outcome will occur induces uncertainty over which numerical value the random variable will take.
- The randomness in the example of X being whether the person voted or not comes from the randomness of that outcome, not in the mapping of "vote" into 1 and "didn't vote" into 0.

DISCRETE RANDOM VARIABLES

- A **discrete random variable** is one that takes on only a finite or countably infinite number of values.

- Countably infinite just means that it takes on any integer and there's no (obvious) upper bound to the values that it can take.
- The most obvious discrete r.v. is the binary r.v., which can only take on two values: 0 and 1.

Examples

- Number of Democrats who win election in the Senate
- An indicator of whether two countries go to war
- The number of times a particular word is used in a document

CONTINUOUS RANDOM VARIABLES

- A **continuous random variable** is one that can take on any real value.
- These are variables for which there are an uncountably infinite number of possible realizations.
- Note that the variables are only approximately continuous—that is, they have a very large number of possible realizations and treating it as continuous is a good approximation.

Examples

- The length of time between two governments in a parliamentary system
- The proportion of voters who turned out
- Budgets allocations to various government programs

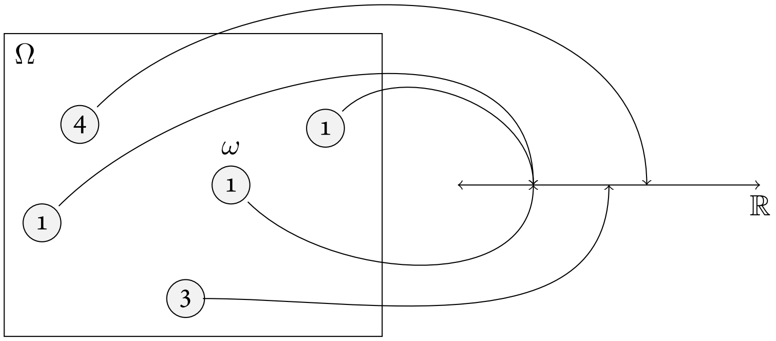
PROBABILITY DISTRIBUTIONS OF RANDOM VARIABLES

- We want some way of compactly representing the information about how likely various outcomes of X are. That is, we need to figure out what the distribution of X will look like.
- The distribution of the random variable is extremely important to our endeavors in statistics. Keep in mind that the probability distribution tells us what kind of data we should expect to see in the world. Our goal in statistics is going to be to learn about the probability distribution of some variable from data that we have.

Where do the probability distributions come from?

- The probabilities associated with each realization of the r.v. come from the underlying experiment and sample space. In this diagram, we can see this process.

We have individual events in the sample space. The r.v. has given these numerical values and plops them over into the real line.



- The probability of the r.v. equaling some number is just the probability that the underlying event occurs. If we were drawing these balls from the sample space with equal probability, then we would find that the probability of drawing a 1 would be 3/5 and the probability of drawing a 3 would be 1/5 and the probability of drawing a 5 would be 0.

- Thus, we can infer the relative distribution of the realizations of the random variable from the relative frequencies/probabilities of the underlying events. Different types of experiments will lead to different types of distributions. Many of these are common across applications, so we can talk about common distributions without explicit reference to the underlying experiment.

- Remember that probabilities on the sample space come from a **data generating process** (DGP)—assumptions about the physical or social world. Assuming that we have independent coin flips induces independent probabilities of 0.5 for each coin flip. Random sampling from a set induces equal probabilities of each object.

- The DGP, then, will also induce the probability distribution for the random variable.

Cumulative distribution functions

- For both discrete and continuous variables, one useful function is the **cumulative distribution function** (cdf), which tells us what the probability is that a variable is less than a particular realization:

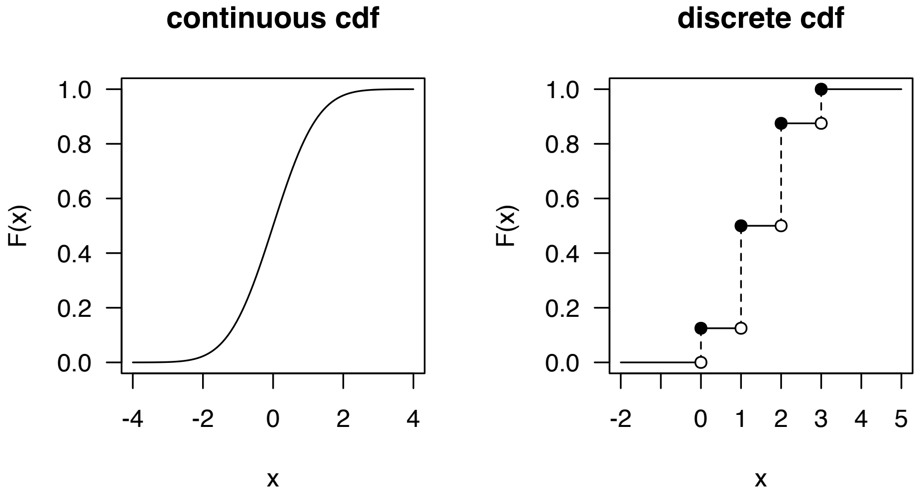
$$F_X(x) \equiv \mathbb{P}(X \leq x).$$

- This function completely describes the distribution of a random variable.
- This general definition is the same for both discrete and continuous variables, but each has a more specific way of writing the cdf.
- A couple of properties: 1) never decreases, 2) limits to 0 toward negative infinity, limits to 1 toward positive infinity, 3) right-continuous (no jumps when we approach a point from the right)
- Note that if we knew the cdf then we could calculate the probability of other events, such as:

  $$\mathbb{P}(a \leq X \leq b) = F(b) - F(a)$$

  You can see that this works by using the following equality which we can prove using the properties of probabilities from last week (noting that the complement of $X \leq a$ is $X \geq a$):

  $$\mathbb{P}(X \leq b) = \mathbb{P}(X \leq a) + \mathbb{P}(X \geq a \cap X \leq b)$$



Discrete r.v.s - probability mass function

- For a discrete r.v., each possible outcome has an associated probability of occurring. Go back to our example on voting. For a particular individual, they have some probability of voting $\mathbb{P}(X = 1)$ and some probability of not voting $\mathbb{P}(X = 0)$.
- But this can generalize as well: we can list out all possible values of the discrete r.v. and also their associated probabilities. This provides a nice summary of the entire distribution of the variable.

- **Probability mass function**: for a discrete random variable, X, we can define the probability mass function (pmf) as:

$$f_X(x_j) = \mathbb{P}(X = x_j) \qquad j = 1, 2, \dots, k$$

- Given this, we can write the cdf of a discrete r.v. is:

$$F_X(x) = \sum_{x_k \leq x} f_X(x_k)$$

- Some properties of the pmf fall out of the properties of probability: $0 \leq f_X(x) \leq 1$ and $\sum_{i=1}^{k} f_X(x_j) = 1$.

Example - random assignment to treatment

- Suppose that we're running a randomized control trial to see if some intervention works—maybe we're random assigning some people to receive GOTV mailer or randomly assigning them to watch negative versus positive ads. Let's say that we did a poor job at recruiting subjects so we only have 3 subjects.

- Here's our procedure for randomly assignment: flip a coin for each unit independently and assign those with heads to Treatment and those with tails to Control. We'll define X to be the number of treated units:

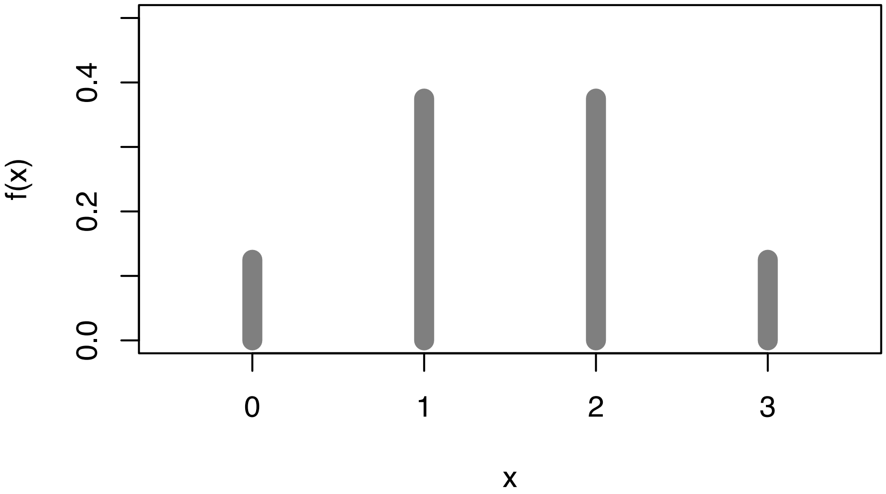
$$X = \begin{cases} 0 & \text{if } (C, C, C) \\ 1 & \text{if } (T, C, C) \text{ or } (C, T, C) \text{ or } (C, C, T) \\ 2 & \text{if } (T, T, C) \text{ or } (C, T, T) \text{ or } (T, C, T) \\ 3 & \text{if } (T, T, T) \end{cases}$$

- We can use the underlying probabilities of the coin flips to calculate the probability of each outcome. First note that $\mathbb{P}(C, T, C) = \mathbb{P}(C)\mathbb{P}(T)\mathbb{P}(C) = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8}$, where the first equality holds by independence of the coin flips and the second by the fair coin assumption. Also, note that this is true for any of the outcomes. Thus,

$$\begin{aligned} f_X(0) &= \mathbb{P}(X = 0) = \mathbb{P}(C, C, C) = \frac{1}{8} \\ f_X(1) &= \mathbb{P}(X = 1) = \mathbb{P}(T, C, C) + \mathbb{P}(C, T, C) + \mathbb{P}(C, C, T) = \frac{3}{8} \\ f_X(2) &= \mathbb{P}(X = 2) = \mathbb{P}(T, T, C) + \mathbb{P}(C, T, T) + \mathbb{P}(T, C, T) = \frac{3}{8} \\ f_X(3) &= \mathbb{P}(X = 3) = \mathbb{P}(T, T, T) = \frac{1}{8} \end{aligned}$$

- What's $\mathbb{P}(X = 4)$?
- We could plot this pmf using R:

```
## plot the pdf of a Normal random variable
plot(x = c(0,1,2,3), y = c(1/8, 3/8, 3/8, 1/8), type = ”h”, ylab = ”f(x)”,
     xlab = ”x”, lwd = 10, xlim = c(-0.5,3.5), ylim = c(0, 0.5), col = ”grey50”)
```



- **Question**: Does this seem like a good way to assign treatment? What is one major problem with it? Can you think of another way to assign treatment that would avoid these problems?

Some important discrete distributions

- Let X be a binary variable with $\mathbb{P}(X = 1) = p$ and, thus, $\mathbb{P}(X = 0) = 1 - p$, where $p \in [0, 1]$. Then we say that X follows a **Bernoulli distribution** with the following pmf:

$$f_X(x) = p^x(1-p)^{1-x} \qquad \text{for } x \in \{0, 1\}.$$

- Probably the most famous distribution for a discrete r.v. is the **discrete uniform distribution** that puts equal probability on each value that X can take:

$$f_X(x) = \begin{cases} 1/k & \text{for } x = 1, \dots, k \\ 0 & \text{otherwise} \end{cases}$$

- Note that we can summarize these distributions with one number—with the discrete distribution it's the number of possible outcomes and with the Bernoulli distribution it is probability of variable being 1.

Continuous r.v.s - probability density function

- With a continuous r.v., we want to do something similar—describe how likely some set of outcomes are. We might think to take the same approach as with a discrete r.v. and just go through each possible value of X and list out its corresponding probability. This approach breaks down, though, when the number of possible values is uncountable because the number of possible realizations is massive (there is an infinite number of them in any subset of the real line).
- This means that we have to take the probability of any particular realization (for example, $2.32879873\ldots$) as 0 and instead we will work with the probability of X being in some set B.
- Thus, we will define the **probability density function** (pdf) for a continuous random variable, X as the function $f_X(x) \geq 0$ for all x and for a set $B \subset \mathbb{R}$:

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx.$$

- The **uniform distribution** on the (a, b) interval is the distribution where the probability of an interval is equal to one over its length. We write $X \sim$ Unif(a, b) and it has the pdf:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

- Note that two conditions on the pdf must be met. First, $f_X(x) \geq 0$ for all x. Second, the total density must be equal to 1. That is,

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

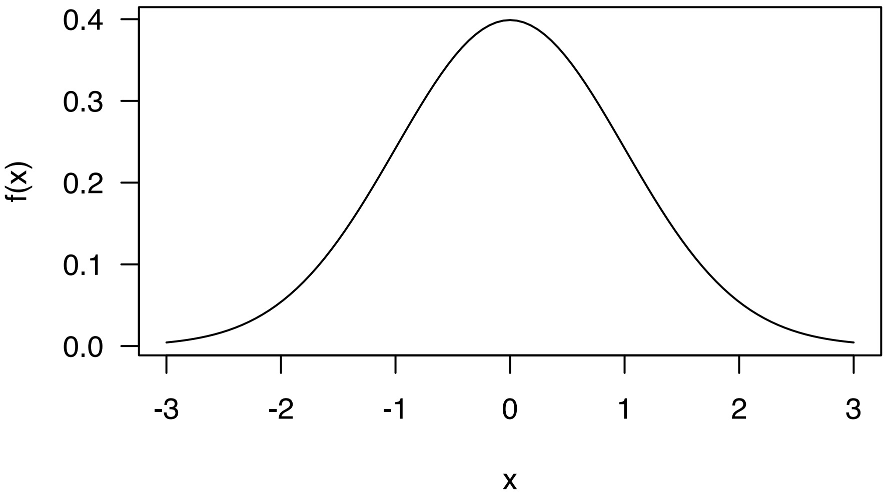
.

- In particular, we have the following, when $a \leq b$, then we can find the probability of X being between a and b as:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

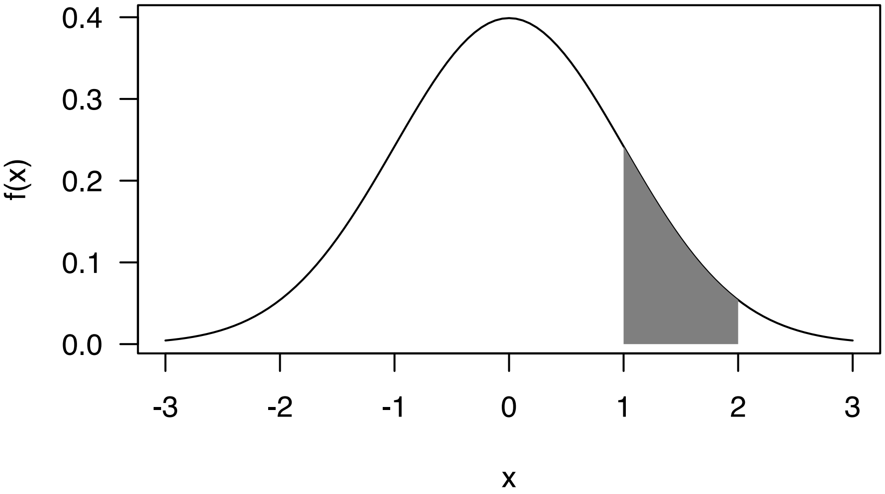
- The pdf gives us information about how likely various outcomes are. Regions with higher values of the pdf are areas where we are more likely to see a realization of X.

```
## plot the pdf of a Normal random variable
curve(dnorm(x), from = -3, to = 3, ylab = ”f(x)”, yaxt = ”n”)
axis(side = 2, las = 2)
```



- **But be careful!** The height of the curve here, $f_X(x)$ is not equal to the probability of x occurring—remember that is 0 for a continuous variable. To get the probability that X will fall in some region, we need to take the integral, which corresponds to the area under the pdf curve:

```
## plot the pdf of a Normal random variable
curve(dnorm(x), from = -3, to = 3, ylab = ”f(x)”, yaxt = ”n”)
axis(side = 2, las = 2)
polygon(c(1,seq(1,2,0.01),2), c(0,dnorm(seq(1,2,0.01)), 0), col = ”grey50”, border = NA)
```



- Given this, we can write the cdf of a continuous r.v. as:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

  Note that here we use t in the place of where x usually is so that it doesn't get mixed up with the x in the integral operation.

- The **Normal distribution** is the classic "bell-shaped" curve. It is extremely useful and ubiquitous in statistics. If X has a Normal distribution, we write $X \sim N(\mu, \sigma^2)$, where μ is the expected value of the distribution and σ^2 is the variance. We'll define those concepts in just a bit.

- The pdf for the Normal distribution is:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

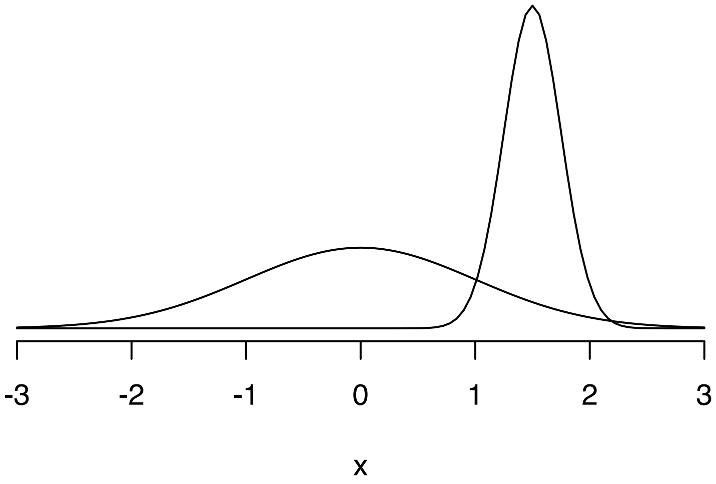
- When the mean is 0 and the variance is 1, we call this the **standard Normal distribution**.

- The reason this distribution comes up so much is that many things follow an approximately Normal distribution.

- We'll cover more distributions as we need them for statistical inference. In Gov 2001, you'll learn about quite a few more distributions.

PROPERTIES OF DISTRIBUTIONS

- So now we have these functions that describe what realizations of a random variable are more likely than others. In general, we're not going to know these distributions, but we often want to learn about them.
- **Question**: What is the difference between these two density curves? How might we summarize this difference?



- As we've seen, distributions can be these complicated mathematical expressions that are hard to interpret. These two distribution have many differences: one is the probability that each places around -1, another is how much they place around 2, and so forth.
- It would be nice, however, to be able to summarize these distributions quickly so that we can have an intuitive understanding of where most of the data will be. To do this, we can think of two features of distribution that are easily quantifiable and informative: the center of the distribution and the spread of that distribution around its center.

MEASURES OF CENTRAL TENDENCY - EXPECTED VALUE

- The central tendency of the distribution is a measure of the "middle" of the distribution. There are a couple of ways we might think about where the middle is.
- These measures are one-number summaries of the distribution in the sense that they represent our best guess of the value of X before we see it. = The measure of central tendency we will focus on in this class is the **expected value**, which is also called the **expectation** or the **mean** of the distribution.

- We refer to expected value of X as $\mathbb{E}[X]$ or μ.

Motivation - calculating averages

- Imagine we had a bunch of numbers an you wanted to calculate the average: (1,1,3,4,4,5). Obviously, you would add them up and divide by the number of items in the sum:
$$\frac{1+1+3+4+4+5}{6} = \frac{18}{6}3$$
- Now, you could always have calculated that a slightly different way:
$$\frac{1}{6} \times 1 + \frac{1}{6} \times 1 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5$$
and if we group terms we get:
$$\frac{2}{6} \times 1 + \frac{1}{6} \times 3 + \frac{2}{6} \times 4 + \frac{1}{6} \times 5$$
- This last expression is another way to calculate the mean: sum up the values in the set, weighted by their proportion in the set. This form is the exact way that we'll think of the mean.

Definition

- As with the distribution, we calculate the expected value differently for discrete and continuous random variables. For both of them, the expected value is a **weighted average** of the realizations weighted by the probability of occurring.
- For discrete X, this is straightforward:
$$\mathbb{E}[X] = \sum_{j=1}^{k} x_j f(x_j)$$
- For continuous X, this is slightly more complicated because we have to use the integral:
$$\mathbb{E}[X] = \int x f(x) dx$$
- **Exercise**: Let X be a Bernoulli r.v. with $\mathbb{P}(X = 1) = p$. Use the definition of the expected value to calculate $\mathbb{E}[X]$.
- Sometimes we will calculate the expectation from the distribution (like in the next example). If we know the distribution of the data, though (Bernoulli, Uniform, Normal), then the expected value is usually just a known function of the parameters (like in the above example with the Bernoulli).

Example - number of treated units

- Let's go back to the number of treated units to figure how many units we should expect to be treated in our experiment:

$$
\begin{aligned}
E[X] = \sum_{j=1}^{k} x_j f(x_j) &= 0 \times f_X(0) + 1 \times f_X(1) + 2 \times f_X(2) + 3 \times f_X(3) \\
&= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times 18 \\
&= 0 + \frac{3}{8} + \frac{6}{8} + \frac{3}{8} = \frac{12}{8} = 1.5
\end{aligned}
$$

- If we look back at the pmf of this distribution, it makes a lot of sense that the answer would 1.5 since that is in the middle of the distribution.
- This answer brings up an interesting feature of the expected value: it doesn't have to be one of the values that the r.v. can take.

Properties of the expected value

- The expected value has a lot of nice properties that make it easy to work with.
- Both of the key properties of expected values are that they are linear. What does that mean?
- **Additivity:** (expectation of sums are sums of expectations)

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

- **Homoegeneity:** Suppose that a and c are constants. Then,

$$\mathbb{E}[aX + c] = a\mathbb{E}[X] + c$$

- **Law of the Unconscious Statistician**, or LOTUS. If $g(X)$ is a function of a discrete random variable, then

$$E[g(X)] = \sum_x g(x) f_X(x),$$

  which basically says that the expected value of the transformation of the random variable is just the weighted average of the transformed outcomes.

Example: list experiments

- Let's say that Y is the number of items that people say upset them with the additional "black family" item and X be the number of items that upset them with just the 3 baseline items. Then, we could write $Y = X + A$, where $A = 1$ if the black neighbors question upset them and $A = 0$ it did not.

- Then, we know that $E[Y] - E[X] = E[A]$, but can you prove that?
- If A is a Bernoulli r.v., then how can we interpret $E[A]$?

MEASURES OF SPREAD

- Now we have some sense of where the middle of the distribution is, but we also want to know how spread out the distribution is around that middle.
- We'll talk about two of these that are closely related: the variance and the standard deviation.

Variance

- The **variance** is the average of the squared distances from the mean. We sometimes denote it σ_X^2. In general, we write it as:

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Since the squared distances are always nonnegative, the variance is also always nonnegative.
- If most of the observations are close to the expected value, then the variance will be closer to 0. If the observations are far from the expected value, then the variance will be higher.
- We can use LOTUS from above to calculate the variance for a discrete random variable:

$$\text{Var}[X] = \sum_x (x - \mathbb{E}[X])^2 f_X(x)$$

- And we can apply the same principle for continuous random variables:

$$\text{Var}[X] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f_X(x) dx$$

- The **standard deviation** is just the (positive) square root of the variance: $\sigma_X = \sqrt{\text{Var}[X]}$. You can interpret this as the average distance from the expected value of the distribution. If someone told you the mean and asked you how far away a random draw from the distribution would be

Example - number of treated units

- Let's go back to the number of treated units to figure out the variance of the number of treated units:

$$\begin{aligned}
E[X] &= \sum_{j=1}^{k}(x_j - E[X])^2 f(x_j) \\
&= (0-1.5)^2 \times f_X(0) + (1-1.5)^2 \times f_X(1) + (2-1.5)^2 \times f_X(2) + (3-1.5)^2 \times f_X(3) \\
&= (-1.5)^2 \times \frac{1}{8} + (-0.5)^2 \times \frac{3}{8} + 0.5^2 \times \frac{3}{8} + 1.5^2 \times 18 \\
&= 2.25 \times \frac{1}{8} + 0.25 \times \frac{3}{8} + 0.25 \times \frac{3}{8} + 2.25 \times 18 = 0.75
\end{aligned}$$

- **Exercise**: What's the standard deviation for this distribution?

Properties of variances

- Variances have slightly different properties than expectations, but there are similar flavors. First, note that the variance of a constant, b, is 0: Var$[b] = 0$. You should use the definition of the variance to convince yourself why that is the case (hint: what's the expected value of a constant?).
- Is the variance linear like the expectation? No, but here's what we can say about the variance:

$$\text{Var}[aX + b] = a^2\text{Var}[X]$$

- We'll say more about the variance of the sum of two random variables Var$(X + Y)$ when we get to multiple random variables next time.