The Binomial Distribution

Patrick Breheny

September 30
So far, we have discussed the probability of single events.

In research, however, the data we collect consists of many events (for each subject, does he/she contract polio?)

We then summarize those events with a number (out of the 400,000 people who got the vaccine, how many contracted polio?)

Such a number is an example of a random variable.
In our sample, we observe a certain value of a random variable. In order to assess the variability of that value, we need to know the chances that our random variable could have taken on different values depending on the true values of the population parameters. This is called a *distribution*. A distribution describes the probability that a random variable will take on a specific value or fall within a specific range of values.
# Examples

<table>
<thead>
<tr>
<th>Random variable</th>
<th>Possible outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td># of copies of a genetic mutation</td>
<td>0, 1, 2</td>
</tr>
<tr>
<td># of children a woman will have in her lifetime</td>
<td>0, 1, 2, ...</td>
</tr>
<tr>
<td># of people in a sample who contract polio</td>
<td>0, 1, 2, ..., n</td>
</tr>
</tbody>
</table>
Listing the ways

- When trying to figure out the probability of something, it is sometimes very helpful to list all the different ways that the random process can turn out.
- If all the ways are equally likely, then each one has probability \( \frac{1}{n} \), where \( n \) is the total number of ways.
- Thus, the probability of the event is the number of ways it can happen divided by \( n \).
For example, the possible outcomes of an individual inheriting cystic fibrosis genes are

\[ CC \quad Cc \quad cC \quad cc \]

If all these possibilities are equally likely (as they would be if the individual’s parents had one copy of each version of the gene), then the probability of having one copy of each version is 2/4
Another example where the outcomes are equally likely is flips of a coin

Suppose we flip a coin three times; what is the probability that exactly one of the flips was heads?

Possible outcomes:

\[
\begin{align*}
HHH & \quad HHT & \quad HTH & \quad HTT \\
TTH & \quad THT & \quad TTH & \quad TTT
\end{align*}
\]

The probability is therefore $3/8$
The binomial coefficients

- Counting the number of ways something can happen quickly becomes a hassle (imagine listing the outcomes involved in flipping a coin 100 times)
- Luckily, mathematicians long ago discovered that when there are two possible outcomes that occur/don’t occur \( n \) times, the number of ways of one event occurring \( k \) times is

\[
\binom{n}{k} = \frac{n!}{k!(n-k)!}
\]

- The notation \( n! \) means to multiply \( n \) by all the positive numbers that come before it (e.g. \( 3! = 3 \cdot 2 \cdot 1 \))
- Note: \( 0! = 1 \)
Calculating the binomial coefficients

- For the coin example, we could have used the binomial coefficients instead of listing all the ways the flips could happen:

\[
\frac{3!}{1!(3-1)!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 1} = 3
\]

- Many calculators and computer programs (including SAS) have specific functions for calculating binomial coefficients, which we will explore in lab.
When sequences are not equally likely

- Suppose we draw 3 balls, with replacement, from an urn that contains 10 balls: 2 red balls and 8 green balls.
- What is the probability that we will draw two red balls?
- As before, there are three possible sequences: $RRG$, $RGR$, and $GRR$, but the sequences no longer have probability $\frac{1}{8}$. 

The probability of each sequence is

\[
\frac{2}{10} \cdot \frac{2}{10} \cdot \frac{8}{10} = \frac{2}{10} \cdot \frac{8}{10} \cdot \frac{2}{10} = \frac{8}{10} \cdot \frac{2}{10} \cdot \frac{2}{10} \approx .03
\]

Thus, the probability of drawing two red balls is

\[
3 \cdot \frac{2}{10} \cdot \frac{2}{10} \cdot \frac{8}{10} = 9.6\%
\]
This line of reasoning can be summarized in the following formula: the probability that an event will occur \( k \) times out of \( n \) is

\[
\frac{n!}{k!(n - k)!} p^k (1 - p)^{n-k}
\]

In this formula, \( n \) is the number of trials, \( p \) is the probability that the event will occur on any particular trial.

We can then use the above formula to figure out the probability that the event will occur \( k \) times.
Example

- According to the CDC, 22% of the adults in the United States smoke.
- Suppose we sample 10 people; what is the probability that 5 of them will smoke?
- We can use the binomial formula, with

\[
\frac{10!}{5!(10-5)!} \cdot 0.22^5 (1-0.22)^{10-5} = 3.7%
\]
What is the probability that our sample will contain two or fewer smokers?

We can add up probabilities from the binomial distribution:

\[ P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2) \]

\[ = 0.083 + 0.235 + 0.298 \]

\[ = 61.7\% \]
The binomial formula – when to use

This formula works for any random variable that counts the number of times an event occurs out of \( n \) trials, provided that the following assumptions are met:

- The number of trials \( n \) must be fixed in advance
- The probability that the event occurs, \( p \), must be the same from trial to trial
- The trials must be independent

If these assumptions are met, the random variable is said to follow a \textit{binomial distribution}, or to be \textit{binomially distributed}
One-sample categorical data

- The binomial distribution plays a central role in the analysis of one-sample categorical data.
- For example, a study at Johns Hopkins estimated the survival chances of infants born prematurely by surveying the records of all premature babies born at their hospital in a three-year period.
- In their study, they found 39 babies who were born at 25 weeks gestation, 31 of which survived at least 6 months.
This type of study has one sample of 39 babies. If some of these babies had received one type of therapy and the rest a different kind of therapy, and we were interested in comparing the two therapies, then we would have two samples. The outcome of this study is categorical, in that a baby either survived for 6 months or it didn’t. If we had instead decided to measure lung function or weight or some continuous measure of health, we would have continuous data. As we will see, recognizing how many samples there are, and what kind of data the outcome is, plays a central role in the proper way to analyze that study.
Generalization to the population

- The Johns Hopkins study observed that $\frac{31}{39} = 79.5\%$ of babies survive after being born at 25 weeks gestation.

- The goal of the study was not to audit their hospital’s performance, but to estimate the percent of babies in other (comparable) hospitals, in future years (although maybe not too far in the future), that would survive early labor.

- This is the generalization they want to make, but how accurate is their percentage?

- Could the actual percent of babies who would survive such an early labor (in other hospitals, in future years) be as high as 95%? As low as 50%?
Confidence interval

- The number of infants who survive will follow a binomial distribution
- Let \( p \) denote the probability that an infant will survive, let \( p_0 \) denote the true, unknown value of that probability, and let \( \hat{p} = .795 \) equal our estimate of that probability based on our sample (this is common notation in statistics to distinguish parameters from estimates)
- In order to build a 95% confidence interval, we need a way to calculate two numbers, \( (p_L, p_U) \) that have a 95% probability of containing \( p_0 \)
- The most natural way of doing this is to find \( p_L \) so that we only have a 2.5% probability of getting \( \hat{p} \) or higher, and \( p_U \) so that there is only a 2.5% probability of obtaining \( \hat{p} \) or lower
Trial and error

- Let’s start by trying out the value $p_L = .5$
- If $p_0 = .5$, what is the probability that 31 or more babies (out of 39) would survive?
- Letting $X$ denote the number of babies who survive,

\[
P(X \geq 31) = P(X = 31) + P(X = 32) + \ldots
\]

\[
= \frac{39!}{31!8!} \cdot .5^{31} (1 - .5)^8 + \ldots
\]

\[
= .000112 + .000028 + \ldots
\]

\[
= .00015
\]

- This is much lower than the 2.5% we were shooting for; we need to raise $p_L$
Finding $p_L$ and $p_U$

This sort of trial and error is tedious to do by hand, but trivial for a computer:
Thus, our confidence interval for the (population) percentage of infants who survive after being born at 25 weeks is (63.5%, 90.7%).

In their study, the Johns Hopkins researchers also found 29 infants born at 22 weeks gestation, none of which survived 6 months.

Applying the same procedure, we obtain the following confidence interval for the percentage of infants who survive after being born at 22 weeks: (0%, 11.9%).
It is relatively rare to have specific hypotheses in one-sample studies.

One very important exception is the collection of paired samples.

In a paired sampling design, we collect \( n \) pairs of observations and analyze the difference between the pairs.
Hypothetical example: A sunblock study

- Suppose we are conducting a study investigating whether sunblock A is better than sunblock B at preventing sunburns.
- The first design that comes to mind is probably to randomly assign sunblock A to one group and sunblock B to a different group.
- This is nothing wrong with this design, but we can do better.
Generally speaking, our ability to make generalizations about the population depends on two factors: *signal* and *noise*

*Signal* is the magnitude of the difference between the two groups – in the present context, how much better one sunblock is than the other.

*Noise* is the variability present in the outcome from all other sources besides the one you’re interested in – in the sunblock experiment, this would include factors like how sunny the day was, how much time the person spent outside, how easily the person burns, etc.

Hypothesis tests depend on the ratio of signal to noise – how easily we can distinguish the treatment effect from all other sources of variability.
To get a larger signal-to-noise ratio, we must either increase the signal or reduce the variability.

The signal is usually determined by nature and out of our control.

Instead, we are going to have to reduce the variability/noise.

If our sunblock experiment were controlled, we could attempt such steps as forcing all participants to spend an equal amount of time outside, on the same day, in an equally sunny area, etc.
Person-to-person variability

- But what can be done about person-to-person variability (how easily certain people burn)?
- A powerful technique for reducing person-to-person variability is *pairing*.  
  - For each person, we can apply sunblock A to one of their arms, and sunblock B to the other arm, and as an outcome, look at the difference between the two arms.
- In this experiment, the items that we randomly sample from the population are pairs of arms belonging to the same person.
Benefits of paired designs

- What do we gain from this?
- As variability goes down,
  - Confidence intervals become narrower
  - Hypothesis tests become more powerful
- How much narrower? How much more powerful?
- This depends on the fraction of the total variability that comes from person-to-person variability
Investigators have come up with all kinds of clever ways to use pairing to cut down on variability:

- Crossover studies
- Family studies
- Split-plot experiments
Pairing in observational studies

- Pairing is also widely used in observational studies
  - Twin studies
  - Matched studies
- In a matched study, the investigator will pair up (“match”) subjects on the basis of variables such as age, sex, or race, then analyze the difference between the pairs
- In addition to increasing power, pairing in observational studies also eliminates (some of the) potential confounding variables
Cystic fibrosis experiment

- As an example of a paired study, we will look at a crossover study of the drug amiloride as a therapy for patients with cystic fibrosis
- Cystic fibrosis is a fatal genetic disease that affects the lungs
- Forced vital capacity (FVC) is the volume of air that a person can expel from the lungs in 6 seconds
- FVC is a measure of lung function, and is often used as a marker of the progression of cystic fibrosis
There were 14 people who participated in the study.

Each participant in the trial received both the drug and the placebo (at different times), “crossing over” to receive the other treatment halfway through the trial.

Like all well-designed crossover trials, the therapy (treatment/placebo) that each participant received first was chosen at random.

Furthermore, there was a washout period during the crossover between the two drug periods.
The outcome

- To determine an outcome, the FVC of the patients was measured at the beginning of each treatment period, and again at the end.
- The outcome is the reduction in lung function over the treatment period.
- So, for example, if a patient’s FVC was 900 at the beginning of the drug period and 850 at the end, the reduction is 50.
- In the actual study, 11 of the 14 patients did better on the drug than on the placebo.
- A hypothesis test informs us whether or not this kind of result could be due to chance alone.
The null hypothesis

- The null hypothesis here is that the drug provides no benefit – that whether the patient received drug or placebo has no impact on their lung function.
- Under the null hypothesis, then, the probability that a patient does better on drug than placebo \( p \) is 50%.
- So the null hypothesis is that \( p_0 = .5 \).
- Essentially, under the null, whether a patient does better on one treatment or another is like flipping a coin.
One way to test this null hypothesis would be to flip a coin 14 times, count the number of heads, and repeat this over and over again to see how unusual "11 heads" is. However, this is unnecessary, as we already have the binomial distribution to calculate these probabilities for us. Under the null hypothesis, the number of patients who do better on the drug than placebo (X) will follow a binomial distribution with \( n = 14 \) and \( p = 0.5 \). This approach to hypothesis testing goes by several names, and could be called the *exact test*, the *binomial test*, or the *sign test*. What we need to do is calculate the \( p \)-value: the probability of obtaining results as extreme or more extreme than the one observed in the data, given that the null hypothesis is true.
“As extreme or more extreme”

- The result observed in the data was that 11 patients did better on the drug.
- But what exactly is meant by “as extreme or more extreme” than 11?
- It is uncontroversial that 11, 12, 13, and 14 are as extreme or more extreme than 11.
- But what about 0? Is that more extreme than 11?
- Under the null, $P(11) = 2.2\%$, while $P(0) = .006\%$.
- So 0 is more extreme than 11, but in a different direction.
One-sided vs. two-sided tests

- Potentially, then, we have two different approaches to calculating this $p$-value:
  - Find the probability that $x \geq 11$
  - Find the probability that $x \geq 11 \cup x \leq 3$ (the number that is as far away from the expected value of 7 as 11 is, but in the other direction)

- These are both reasonable things to do, and intelligent people have argued both sides of the debate

- However, the statistical and scientific community has for the most part come down in favor of the latter – the so called “two-sided test”

- For this class, all of our tests will be two-sided tests
Thus, the $p$-value of the sign test is

$$p = P(x \leq 3) + P(x \geq 11)
= P(x = 0) + \cdots + P(x = 3) + P(x = 11) + \cdots + P(x = 14)
= 0.006\% + 0.09\% + 0.6\% + 2.2\% + 2.2\% + 0.6\% + 0.09\% + 0.006\%
= 5.7\%$$

Seeing 11 out of 14 patients do better on one treatment than another is therefore reasonably unlikely

This is moderate evidence against the null hypothesis