The Binomial Distribution

Patrick Breheny

September 17

Outcomes and summary statistics

- So far, we have discussed the probability of events
- In most studies, however, it is usually easier to work with a summary statistic than the actual sample space
- For example, in the polio study, the relevant information in the study can be summarized by the number of people who contracted polio; this is vastly easier to think about than all possible outcomes of all possible samples that could be drawn from the population

Random variables

- A numerical summary X of an outcome is called a *random variable*
- More formally, a random variable is a function mapping the sample space S to the real numbers $\mathbb{R}$

Random variable	Possible outcomes
# of copies of a genetic mutation	0,1,2
# of children a woman will have in her lifetime	0,1,2,...
# of people in a sample who contract polio	0,1,2,...,n

Distributions

- Once the random process is complete, we observe a certain value of a random variable
- In order to make inferences, 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

Distribution: technical definition

Definition: Given a random variable X and probability function P defined on a sample space, the *distribution* (or *law*) of X is a function that, for a given interval B[1], gives

$$P(X \in B)$$

[1] The interval may simply be a single point, e.g., $[5, 5]$

Cumulative distribution function

Definition: The *cumulative distribution function* (CDF) of a random variable X is

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

- Note that X is the random variable and x is the (constant) argument of the function
- Note that the distribution uniquely defines the CDF by setting $B = (-\infty, x]$
- Less obviously, the CDF uniquely defines the distribution; for example, $P(X \in [L, U]) = F(U) - \lim_{x \nearrow L} F(x)$

Probability mass function

If X is discrete (i.e., takes on only a finite or countable number of values), we can also describe point probabilities:

Definition: The *probability mass function* (PMF) of a random variable X is given by

$$f(x) = P(X = x)$$

- Again, X is the random variable and x is the argument
- It is easy to see in this case that there is a one-to-one relationship between PMFs and CDFs; for example, $F(x) = \sum_{s \leq x} f(s)$
- A common convention is to use an uppercase letter for a CDF and the lower case letter for its PMF

Probability density functions

If X is continuous (in the sense that $F(x)$ is a continuous function), the PMF is not useful since $f(x) = 0 \, \forall x$; in this case, we need to introduce the concept of probability "density":

Definition: The *probability density function* (PDF) of a random variable X is given by

$$f(x) = \frac{d}{dx} F(x)$$

- Although $P(X = x) = 0$, we can still talk about the density (probability per infinitesimally small area) at a point x
- A similar relationship again holds between PDF and CDF:

$$F(x) = \int_{s \leq x} f(s) ds$$

Listing the ways

- The most straightforward way of figuring out the probability of something is to list all the elements of the sample space
- 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

Coin example

- For example, suppose we flip a coin three times; what is the probability that exactly one of the flips was heads?
- Possible outcomes:

$$\begin{array}{cccc} HHH & HHT & HTH & HTT \\ THH & THT & TTH & TTT \end{array}$$

- The probability is therefore 3/8

The binomial coefficients

- Listing all the elements of the sample space is often impractical, however (imagine listing the outcomes involved in flipping a coin 100 times)
- Luckily, when there are only two possible outcomes, we can apply the following theorem:
- **Binomial theorem:** For a binary process repeated n times, the number of sequences in which one outcome occurs k times is

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

these numbers are known as the *binomial coefficients*

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 $\frac{1}{8}$

When sequences are not equally likely (cont'd)

- Instead, 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

$$3 \cdot \frac{2}{10} \cdot \frac{2}{10} \cdot \frac{8}{10} = 9.6\%$$

The binomial formula

- We can summarize this result into the following formula:
- **Theorem:** Given a sequence of n independent events that occur with probability p, the probability that an event will occur k times is

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

- Letting X denote the number of times the event occurs, the above yields the PMF of X and therefore defines a specific distribution
- This distribution is called the *binomial distribution*, and X is said to "follow a binomial distribution" or to be "binomially distributed"

Example

- According to the CDC, 22% of the adults in the United States smoke
- **Example:** Suppose we sample 10 people; what is the probability that 5 of them will smoke?
- **Example:** Suppose we sample 10 people; what is the probability that 2 or fewer will smoke?

Summary

- Definitions: random variable, distribution, cumulative distribution function, probability mass function, probability density function
- Binomial coefficients:

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

- Binomial distribution:

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