The Binomial Distribution

Patrick Breheny

February 20

Random variables

- 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 200,000 people who got the vaccine, how many contracted polio?)
- Such a number is an example of a *random variable*

Distributions

- 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

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

Coin example

- As we have said, calculating probabilities is as simple as counting when all outcomes are equally likely
- 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

- 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 x times is

$$\frac{n!}{x!(n-x)!}$$

- 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:

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

- Many calculators and computer programs (including R) have specific functions for calculating binomial coefficients:

```
choose(3,1)
# [1] 3
choose(10,2)
# [1] 45
```

Combinations

- Another way of thinking about what the binomial coefficients mean is with respect to combinations
- A *combination* is the selection of items from a set, where order is unimportant
- For example, if we had three fruits, an apple, a banana, and a cherry, and we wished to select two of them, this is exactly the same as before, just a different way of thinking about it

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)

- 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 distribution

- This line of reasoning can be summarized in the following formula, known as the binomial distribution
- **Binomial distribution:** The probability that an event will occur x times out of n trials is

$$\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x},$$

  where π is the probability that the event will occur on any particular trial (here, π is *not* the number 3.14159...)
- For any n and π, the above formula tells us the probability of seeing every single possible outcome: $x = 0, 1, \ldots, n$

Example

- According to the CDC, 22% of the adults in the United States smoke
- Suppose we sample 10 people; what is the probability that 3 of them will smoke?
- We can use the binomial formula, with

$$\frac{10!}{3!(10-3)!}.22^3(1-.22)^{10-3} = 22.4\%$$

- There is also a shortcut formula in R for this:

```
dbinom(3, 10, 0.22)
# [1] 0.2244458
```

Example (cont'd)

- What is the probability that our sample will contain two or fewer smokers?
- We can add up probabilities from the binomial distribution:

$$\begin{aligned} P(X \leq 2) &= P(X=0) + P(X=1) + P(X=2) \\ &= .083 + .235 + .298 \\ &= 61.7\% \end{aligned}$$

- Or, in R:

```
dbinom(0:2, size=10, prob=.22)
# [1] 0.08335776 0.23511163 0.29841091
pbinom(2, size=10, prob=.22)
# [1] 0.6168803
```

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, π, 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 *binomial distribution*, or to be *binomially distributed*

Summary

- A random variable is a number that can equal different values depending on the outcome of a random process
- The distribution of a random variable describes the probability that the random variable will take on those different values
- The number of ways to choose x things out of n possibilities:

$$\frac{n!}{x!(n-x)!}$$

- The probability that an event will occur x times out of n trials, if π is the probability on any given trial:

$$\frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$$