Lecture 7: Geometric & Binomial distributions

Statistics 101
Mine Çetinkaya-Rundel
February 7, 2012

Announcements

Due: HW 2 at the beginning of class on Thursday:
- Clarification on Exercise 3.4 parts (c) and (d): Mary and Leo’s percentile mean the proportion of people whose finishing times are lower than theirs. Note that this doesn’t mean ”% of people they performed better than” because in a triathlon “performing better” means finishing faster.

Recap

Online quiz 2- commonly missed questions

Question 3:
Q3: More than three-quarters of the nation’s colleges and universities now offer online classes, and about 23% of college graduates have taken a course online. 39% of those who have taken a course online believe that online courses provide the same educational value as one taken in person, a view shared by only 27% of those who have not taken an online course. At a coffee shop you overhear a recent college graduate discussing that she doesn’t believe that online courses provide the same educational value as one taken in person. What’s the probability that she has taken an online course before?

<table>
<thead>
<tr>
<th></th>
<th>took online course</th>
<th>didn’t take online course</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>valuable</td>
<td></td>
<td></td>
<td>0.23</td>
</tr>
<tr>
<td>not valuable</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>total</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Review question

Which is the correct notation for the following probability?
“At a coffee shop you overhear a recent college graduate discussing that she doesn’t believe that online courses provide the same educational value as one taken in person. What’s the probability that she has taken an online course before?”

(a) \( P(\text{took online course} \mid \text{not valuable}) \)
(b) \( P(\text{not valuable} \mid \text{took online course}) \)
(c) \( P(\text{not valuable} \text{ and took online course}) \)
(d) \( P(\text{took online course and not valuable}) \)
(e) \( P(\text{valuable} \mid \text{didn’t take online course}) \)
These experiments measured the willingness of study participants to obey an authority figure who instructed them to perform acts that conflicted with their personal conscience.

- Milgram found that about 65% of people would obey authority and give such shocks.
- Over the years, additional research suggested this number is approximately consistent across communities and time.

Each person in Milgram’s experiment can be thought of as a trial.

- A person is labeled a success if she refuses to administer a severe shock, and failure if she administers such shock.
- Since only 35% of people refused to administer a shock, probability of success is $p = 0.35$.
- When an individual trial has only two possible outcomes, it is called a Bernoulli random variable.

Dr. Smith wants to repeat Milgram’s experiments but she only wants to sample people until she finds someone who will not inflict a severe shock. What is the probability that she stops after the first person?

$$P(1^{st} \text{ person refuses}) = 0.35$$

... the third person?

$$P(1^{st} \text{ and } 2^{nd} \text{ shock, } 3^{rd} \text{ refuses}) = \frac{S}{0.65} \times \frac{S}{0.65} \times \frac{R}{0.35} = 0.65^2 \times 0.35 \approx 0.15$$

... the tenth person?

$$P(1^{st} \text{ and } 2^{nd} \text{ shock, } 3^{rd} \text{ refuses}) = \frac{S}{0.65} \times \frac{S}{0.65} \times \frac{R}{0.35} = 0.65^2 \times 0.35 \approx 0.15$$
**Geometric distribution** describes the waiting time until a success for independent and identically distributed (iid) Bernoulli random variables.

- independence: outcomes of trials don’t affect each other
- identical: the probability of success is the same for each trial

### Geometric probabilities

If $p$ represents probability of success, $(1 - p)$ represents probability of failure, and $n$ represents number of independent trials

$$P(\text{success on the } n^{th} \text{ trial}) = (1 - p)^{n-1} p$$

### Expected value

How many people is Dr. Smith expected to test before finding the first one that refuses to administer the shock?

The expected value, or the mean, of a geometric distribution is defined as $\frac{1}{p}$.

$$\mu = \frac{1}{p} = \frac{1}{0.35} = 2.86$$

She is expected to test 2.86 people before finding the first one that refuses to administer the shock. But how can she test a non-whole number of people?

### Expected value and its variability

**Mean and standard deviation of geometric distribution**

$$\mu = \frac{1}{p}, \quad \sigma = \sqrt{\frac{1 - p}{p^2}}$$

- Going back to Dr. Smith’s experiment:

  $$\sigma = \sqrt{\frac{1 - p}{p^2}} = \sqrt{\frac{1 - 0.35}{0.35^2}} = 2.3$$

- Dr. Smith is expected to test 2.86 people before finding the first one that refuses to administer the shock, give or take 2.3 people.
- These values only makes sense in the context of repeating the experiment many many times.
Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let’s call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of “exactly 1 of them refuses to administer the shock”:

**Scenario 1:**  
\[
\frac{0.35}{(A) \text{ refuse}} \times \frac{0.65}{(B) \text{ shock}} \times \frac{0.65}{(C) \text{ shock}} \times \frac{0.65}{(D) \text{ shock}} = 0.0961
\]

**Scenario 2:**  
\[
\frac{0.65}{(A) \text{ shock}} \times \frac{0.35}{(B) \text{ refuse}} \times \frac{0.65}{(C) \text{ shock}} \times \frac{0.65}{(D) \text{ shock}} = 0.0961
\]

**Scenario 3:**  
\[
\frac{0.65}{(A) \text{ shock}} \times \frac{0.65}{(B) \text{ shock}} \times \frac{0.35}{(C) \text{ refuse}} \times \frac{0.65}{(D) \text{ shock}} = 0.0961
\]

**Scenario 4:**  
\[
\frac{0.65}{(A) \text{ shock}} \times \frac{0.65}{(B) \text{ shock}} \times \frac{0.65}{(C) \text{ shock}} \times \frac{0.35}{(D) \text{ refuse}} = 0.0961
\]

The probability of exactly 1 of 4 people refusing to administer the shock is the sum of all of these probabilities.

\[
0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844
\]

**Counting the # of scenarios**

Earlier we wrote out all possible scenarios that fit the condition of exactly one person refusing to administer the shock. If \( n \) was larger and/or \( k \) was different than 1, writing out the scenarios would get even more tedious. For example, what if \( n = 9 \) and \( k = 2 \):

\[
\begin{align*}
RRSSSSSSS \\
SRRSSSSSS \\
SSRRSSSSSS \\
SSRRESSRSS \\
SSSSRSSSSR \\
\end{align*}
\]

Writing out all possible scenarios is incredibly tedious and prone to errors.

**Calculating the # of scenarios**

The question from the prior slide asked for the probability of given number of successes, \( k \), in a given number of trials, \( n \), \((k = 1\) success in \( n = 4 \) trials), and we calculated this probability as

\[
\# \text{ of scenarios} \times P(\text{single scenario})
\]

- \( \# \text{ of scenarios} \): there is a less tedious way to figure this out, we’ll get to that shortly...
- \( P(\text{single scenario}) = p^k (1-p)^{n-k} \)

The **Binomial distribution** describes the probability of having exactly \( k \) successes in \( n \) independent Bernoulli trials with probability of success \( p \).

Note: You can also use R for these calculations:

```r
> choose(9, 2)  
[1] 36
```
Properties of the choose function

- If $k = 1$, only 1 of the $n$ trials result in a success, it could be the first, the second, ···, or the $n^{th}$ trial, so there are $n$ ways this can happen:
  \[
  \binom{n}{1} = n
  \]

- If $k = n$, all $n$ trials result in a success, and there’s only one way this can happen:
  \[
  \binom{n}{n} = 1
  \]

- If $k = 0$, all $n$ trials result in a failure, and there’s only one way this can happen as well:
  \[
  \binom{n}{0} = 1
  \]

Binomial distribution (cont.)

Binomial probabilities

If $p$ represents probability of success, $(1 - p)$ represents probability of failure, $n$ represents number of independent trials, and $k$ represents number of successes

\[
P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} \ p^k \ (1 - p)^{n-k}
\]

We can use the binomial distribution to calculate the probability of $k$ successes in $n$ trials, as long as

- the trials are independent
- the number of trials, $n$, is fixed
- each trial outcome can be classified as a success or a failure
- the probability of success, $p$, is the same for each trial

Clicker question

A January 27, 2012 Gallup survey suggests that 48% of Americans would vote for Obama over Romney if the presidential election was held that day. Among a random sample of 10 Americans what is the probability that exactly 8 would vote for Obama over Romney?

(a) pretty low
(b) pretty high

Clicker question

A January 27, 2012 Gallup survey suggests that 48% of Americans would vote for Obama over Romney if the presidential election was held that day. Among a random sample of 10 Americans what is the probability that exactly 8 would vote for Obama over Romney?

(a) $0.48^8 \times 0.52^2$
(b) $\binom{8}{10} \times 0.48^8 \times 0.52^2$
(c) $\binom{8}{10} \times 0.48^8 \times 0.52^2$
(d) $\binom{10}{8} \times 0.48^2 \times 0.52^8$

**Expected value**

A January 27, 2012 Gallup survey suggests that 48% of Americans would vote for Obama over Romney if the presidential election was held that day. Among a random sample of 100 people, how many would you expect to vote for Obama?

- Easy enough, $100 \times 0.48 = 48$.
- Or more formally, $\mu = np = 100 \times 0.48 = 48$.
- But this doesn’t mean in every random sample of 100 people exactly 48 will vote for Obama. In some samples this value will be less, and in others more. How much would we expect this value to vary?

**Mean and standard deviation of binomial distribution**

$$
\mu = np \quad \sigma = \sqrt{np(1 - p)}
$$

- Going back to the voters:
  $$
  \sigma = \sqrt{np(1 - p)} = \sqrt{100 \times 0.48 \times 0.52} \approx 5
  $$
- We would expect 48 out of 100 randomly sampled voters, give or take 5.

*Note*: Mean and standard deviation of a Binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

**Unusual observations**

Using the notion that *observations that are more than 2 standard deviations away from the mean are considered unusual* and the mean and the standard deviation we just computed, we can calculate a range for how many people we should expect to find in random samples of 100 that are planning to vote for Obama in the 2012 presidential election.

$$
48 \pm 2 \times 5 = (38, 58)
$$

**Clicker question**

A January 2012 Gallup report suggests that 7.5% of 18-29 year old Americans have been diagnosed with high blood pressure. Would a random sample of 1,000 adults in this age range where only 60 of them have high blood pressure be considered unusual?

(a) Yes (b) No

http://www.gallup.com/poll/152108/Key-Chronic-Diseases-Decline.aspx
A recent study found that “Facebook users get more than they give”. For example:

- 40% of Facebook users in our sample made a friend request, but 63% received at least one request.
- Users in our sample pressed the like button next to friends’ content an average of 14 times, but had their content “liked” an average of 20 times.
- Users sent 9 personal messages, but received 12.
- 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo.

Any guesses for how this pattern can be explained?

This study also found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

We are given that \( n = 245 \), \( p = 0.25 \), and we are asked for the probability \( P(K \geq 70) \).

\[
P(X \geq 70) = P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245)
\]

\[
= P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245)
\]

This seems like an awful lot of work...

Correction for the normal approximation

We apply a 0.5 correction in order to account for the probability of exactly 70 “successes.”
Clicker question

What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?

- (a) 0.0984
- (b) 0.1112
- (c) 0.8888
- (d) 0.9016

Low large is large enough?

The sample size is considered large enough if the expected number of successes and failures are both at least 10.

\[ np \geq 10 \quad \text{and} \quad n(1-p) \geq 10 \]

Clicker question

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution? There are two correct answers.

- (a) \( n = 100, p = 0.8 \)
- (b) \( n = 25, p = 0.6 \)
- (c) \( n = 150, p = 0.05 \)
- (d) \( n = 500, p = 0.015 \)