Lecture 8: Geometric and Binomial distributions

Statistics 101

Mine Çetinkaya-Rundel

September 22, 2011

Dangling thread from last time

From OpenIntro quiz 2:

\[
P(\text{took} \mid \text{not valuable}) = \frac{P(\text{took and not valuable})}{P(\text{not valuable})} = \frac{0.1403}{0.1403 + 0.5621} = 0.1997
\]

Clicker question (graded)

Which of the following is false?

(a) The Z score for the mean of a distribution of any shape is 0.
(b) Half the observations in a distribution of any shape have positive Z scores.
(c) The median of a right skewed distribution has a negative Z score.
(d) Majority of the values in a left skewed distribution have positive Z scores.

Milgram experiment

- Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1963.
- Experimenter (E) orders the teacher (T), the subject of the experiment, to give severe electric shocks to a learner (L) each time the learner answers a question incorrectly.
- The learner is actually an actor, and the electric shocks are not real, but a prerecorded sound is played each time the teacher administers an electric shock.
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.

---

Simulation of Milgram’s experiment

Imagine a hat with 100 pieces of paper in it, 35 are marked “refuse” and 65 are marked “shock”. These represent the subjects in Milgram’s experiment.

```r
> outcomes <- c(rep("refuse", 35), rep("shock", 65))
```

Let’s simulate a sample of 10 subjects, with replacement so that the probability of success stays constant at 0.35 at each trial.

```r
> simSample <- sample(outcomes, size = 10, replace = TRUE)
```

Now let’s take a look at who is in our simulated sample:

```r
> table(simSample)
simSample         shock  refuse
           7           3
```

The sample proportion of successes, \( \hat{p} \), is \( \frac{3}{10} \).

---

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?

\[
P(9 \text{ shock, } 10^{th} \text{ refuses}) = \frac{S}{0.65} \times \cdots \times \frac{S}{0.65} \times \frac{R}{0.35} \left[ \frac{9 \text{ of these}}{0.65^9 \times 0.35} \right] = 0.65^9 \times 0.35 \approx 0.0072
\]
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

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

#### 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**

\[
\mu = \frac{1}{p} \quad \text{and} \quad \sigma = \sqrt{\frac{1 - p}{p^2}}
\]

- 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.

Clicker question

Can we calculate the probability of rolling a 6 for the first time on the 6\(^{th}\) roll of a die using the geometric distribution?

- (a) no, on the roll of a die there are more than 2 possible outcomes
- (b) yes, why not
Suppose we randomly select four individuals to participate in Dr. Smith’s experiment. What is the probability that exactly one 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 one of them refusing to administer the shock”:

- A = refuse, B = shock, C = shock, D = shock: \(0.35 \times 0.65^3 = 0.0961\)
- A = shock, B = refuse, C = shock, D = shock: \(0.35 \times 0.65^3 = 0.0961\)
- A = shock, B = shock, C = refuse, D = shock: \(0.35 \times 0.65^3 = 0.0961\)
- A = shock, B = shock, C = shock, D = refuse: \(0.35 \times 0.65^3 = 0.0961\)

The probability of exactly one out of four people refusing to administer the shock is the sum of all of these probabilities.

\[4 \times 0.0961 = 0.3844\]

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

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 \(k = 2\):

RRSSSSSSS
SRRSSSSSS
SSRSSSSSS
SSRSSSSSS
SSSSSSSSS

Writing out all possible scenarios is way too tedious, and is prone to errors.

Calculating the # of scenarios

Choose function

The choose function is useful for calculating the number of ways to choose \(k\) successes in \(n\) trials.

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

- \(k = 1, n = 4\): \(\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4 \times 3 \times 2 \times 1}{1 \times 3 \times 2 \times 1} = 4\)
- \(k = 2, n = 9\): \(\binom{9}{2} = \frac{9!}{2!(9-1)!} = \frac{9 \times 8 \times 7!}{2 \times 1 \times 7!} = \frac{72}{2} = 36\)

Note: You can also use R for these calculations:
> 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

### Expected value

A September 2011 Gallup survey suggests that 38% of Americans are planning on voting for Barack Obama in the 2012 presidential election. Among a random sample of 100 people, how many would you expect to vote for Obama?

- Easy enough, $100 \times 0.38 = 38$.
- Or more formally, $\mu = np = 100 \times 0.38 = 38$.
- But this doesn’t mean in every random sample of 100 people exactly 38 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?

$$\mu = np$$

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

- Going back to the voters:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.38 \times (1 - 0.38)} = 4.85$$

- We would expect 38 out of 100 randomly sampled voters, give or take 4.85.
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.

\[ 38 \pm 2 \times 4.85 = (28.3, 47.7) \]

**Clicker question**

A September 21, 2011 Gallup report suggests that about 24.2% of 18-25 year olds in the US are uninsured. Would a random sample of 1,000 18-25 year olds where 220 of them are uninsured be considered unusual?

(a) Yes  
(b) No

**Histograms of number of successes**

Hollow histograms of samples from the binomial model where \( p = 0.10 \) and \( n = 10, 30, 100, \) and 300.

- When the sample size is large enough, the binomial distribution with parameters \( n \) and \( p \) can be approximated by the normal model with parameters \( \mu = np \) and \( \sigma = \sqrt{np(1-p)} \).
- The sample size is considered large enough if the expected number of successes and failures are both at least 10.

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

Try it yourself at [http://www.socr.ucla.edu/htmls/SOCR_Distributions.html](http://www.socr.ucla.edu/htmls/SOCR_Distributions.html).
An example

Approximately 20% of the US population smokes cigarettes. A local government believed their community had a lower smoker rate in their community and commissioned a survey of 400 randomly selected individuals. The survey found that only 70 of the 400 participants smoke cigarettes. If the true proportion of smokers in the community was really 20%, what is the probability of observing 70 or fewer smokers in a sample of 400 people?

We are given that $n = 400$, $p = 0.20$, and we are asked for the $P(k \leq 70)$.

$$P(K \leq 70) = P(K = 0 \text{ or } K = 1 \text{ or } \cdots \text{ or } K = 70)$$

$$= P(K = 0) + P(K = 1) + \cdots + P(K = 70)$$

This seems like an awful lot of work...

Statistics 101 (Mine Çetinkaya-Rundel) | L8: Geometric and Binomial | September 22, 2011 24 / 27

An example (cont.)

Instead we can approximate the binomial distribution using the normal distribution.

$$\mu = np = 400 \times 0.20 = 80$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{400 \times 0.20 \times (1-0.20)} = 8$$

$$Z = \frac{X - \mu}{\sigma} = \frac{70.5 - 80}{8} = -1.19$$

$$P(K \leq 70) = P(X < 70.5) = P(Z < -1.19) = 0.1170$$

Note: We can use the normal approximation because $np = 400 \times 0.20 = 80$ and $n(1-p) = 400 \times 0.80 = 320$ are both greater than 10.

Clicker question

Do these data support the local government’s belief that their community has an unusually low smoker rate compared to the population at large?

(a) Yes (b) No