Unit 2: Probability and distributions
Lecture 2: Binomial distribution

Statistics 101

Mine Çetinkaya-Rundel

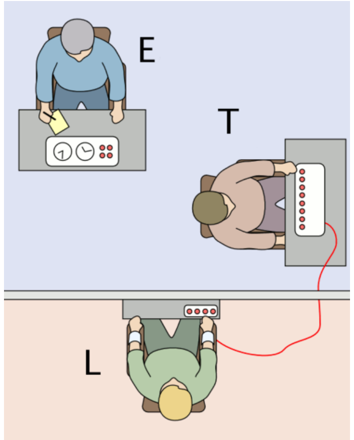
September 19, 2013

Announcements

- Christine's OH moved to Monday *7-9 pm*.
- PA2 available at 5pm today, due 5pm tomorrow (Friday) - scores and feedback will be released later Friday evening, and a review of commonly missed questions will be posted on Piazza over the weekend
- RA3 next Tuesday - Unit 3 materials posted on the website
- PS3 posted (due next Thursday)
- Projects - the Data&GIS Library is open for help, take advantage of it! *http://library.duke.edu/data/about/schedule.html*

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.



Milgram experiment (cont.)

- 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, and only 35% refused.
- Over the years, additional research suggested this number is approximately consistent across communities and time.

Binary outcomes

- 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 also called a Bernoulli random variable.

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: $\underset{\text{(A) refuse}}{\underline{0.35}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

Scenario 2: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) refuse}}{\underline{0.35}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

Scenario 3: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) refuse}}{\underline{0.35}} \times \underset{\text{(D) shock}}{\underline{0.65}} = 0.0961$

Scenario 4: $\underset{\text{(A) shock}}{\underline{0.65}} \times \underset{\text{(B) shock}}{\underline{0.65}} \times \underset{\text{(C) shock}}{\underline{0.65}} \times \underset{\text{(D) refuse}}{\underline{0.35}} = 0.0961$

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

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

Binomial distribution

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

$$\#\ of\ scenarios \times P(single\ scenario)$$

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

  probability of success to the power of number of successes, probability of failure to the power of number of failures

The *Binomial distribution* describes the probability of having exactly k successes in n independent Bernouilli 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, for example, $n = 9$ and $k = 2$:

*RR*SSSSSSS
S*RR*SSSSSS
SS*RR*SSSSS
...
SS*R*SS*R*SSS
...
SSSSSSS*RR*

writing out all possible scenarios would be incredibly tedious and 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.

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

- $k = 1, n = 4$: $\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4\times3\times2\times1}{1\times(3\times2\times1)} = 4$
- $k = 2, n = 9$: $\binom{9}{2} = \frac{9!}{2!(9-1)!} = \frac{9\times8\times7!}{2\times1\times7!} = \frac{72}{2} = 36$

Note: You can also use R for these calculations:

```
> choose(9,2)
[1] 36
```


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

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

Clicker question

Which of the following is not a condition that needs to be met for the binomial distribution to be applicable?

(a) the trials must be independent
(b) the number of trials, n, must be fixed
(c) each trial outcome must be classified as a *success* or a *failure*
(d) the number of desired successes, k, must be greater than the number of trials
(e) the probability of success, p, must be the same for each trial

Clicker question

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) pretty high
(b) pretty low

Gallup: *http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx* , January 23, 2013.

Clicker question

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$
(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$
(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2$
(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$

Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese.

Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, $100 \times 0.262 = 26.2$.
- Or more formally, $\mu = np = 100 \times 0.262 = 26.2$.
- But this doesn't mean in every random sample of 100 people exactly 26.2 will be obese. In fact, that's not even possible. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Expected value and its variability

Mean and standard deviation of binomial distribution

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

- Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

- We would expect 26.2 out of 100 randomly sampled American to be obese, give or take 4.4.

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 the plausible number of obese Americans in random samples of 100.

$$26.2 \pm (2 \times 4.4) = (17.4, 35)$$

Clicker question

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) No

(b) Yes

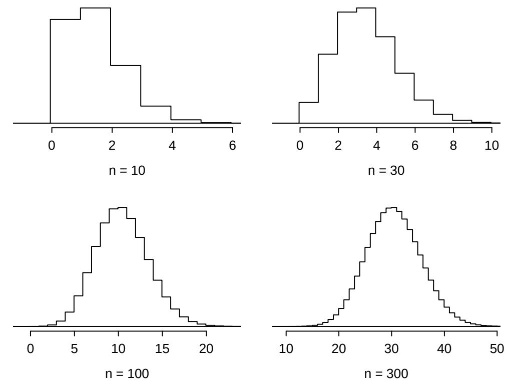
	Excellent	Good	Only fair	Poor	Total excellent/ good
	%	%	%	%	%
Independent private school	31	47	13	2	78
Parochial or church-related schools	21	48	18	5	69
Charter schools	17	43	23	5	60
Home schooling	13	33	30	14	46
Public schools	5	32	42	19	37

Gallup, Aug. 9-12, 2012

http://www.gallup.com/poll/156974/private-schools-top-marks-educating-children.aspx

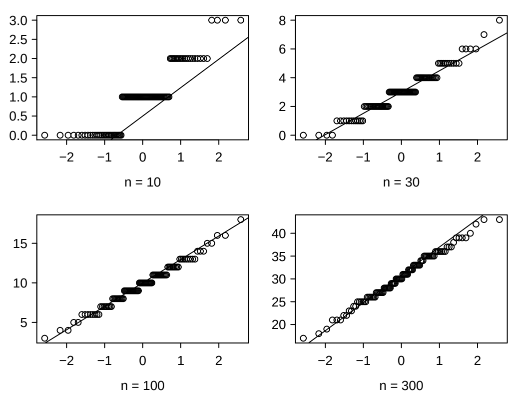
Histograms of number of successes

Hollow histograms of samples from the binomial model where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



Normal probability plots of number of successes

Normal probability plots of samples from the binomial model where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



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.

$$np \geq 10 \qquad \text{and} \qquad n(1-p) \geq 10$$

Clicker question

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

(a) $n = 100, p = 0.95$

(b) $n = 25, p = 0.45$

(c) $n = 150, p = 0.05$

(d) $n = 500, p = 0.015$

An analysis of Facebook users

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?

http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

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 $P(K \geq 70)$.

$$P(X \geq 70) = P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \cdots \text{ or } K = 245)$$
$$= P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245)$$

This seems like an awful lot of work...

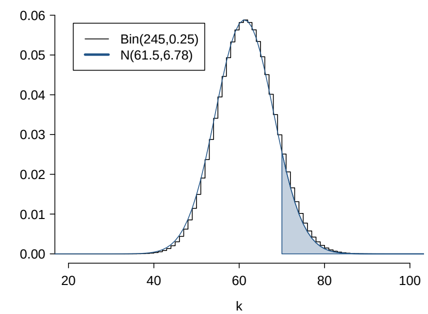
Normal approximation to the binomial

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

- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.



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) 0.0251

(b) 0.0985

(c) 0.1128

(d) 0.9015