Workshop Statistics: Discovery with Data, Second Edition

Topic 16: Sampling Distributions I: Proportions

Activity 16-5: Parameters vs. Statistics

(a) statistic p-hat
(b) parameter theta
(c) statistic x-bar
(d) parameter mu
(e) parameter sigma
(f) statistic s
(g) parameter theta
(h) statistic p-hat
(i) parameter theta
(j) parameter mu
(k) statistic p-hat
(l) statistic x-bar
(m) parameter mu
(n) statistic p-hat
(o) statistic x-bar

Activity 16-6: Presidential Votes

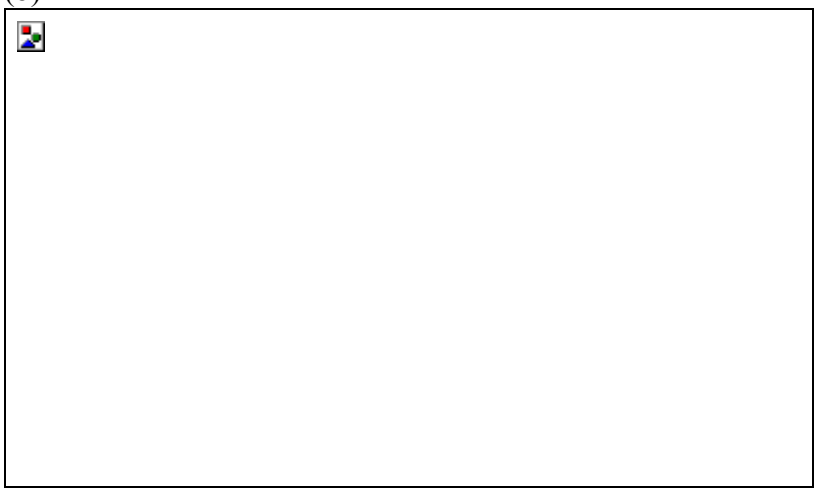
(a) No, sampling variability suggests that these percentages would not hold exactly in the sample of size 100.
(b) No, again due to sampling variability.
(c) This standard deviation would be sqrt(.49*.51/100) = .04999, or about .05.
(d) About 95% of the samples would give a sample proportion of Clinton voters within +/- 2(.05) = .10 of .49, that is between .39 and .59.
(e) This standard deviation would be sqrt(.41*.59/100) = .049.
(f) This standard deviation would be sqrt(.08*.92/100) = .027.
(g) The most variation would occur with the Clinto proportions, the least with the Perot proportions.

Activity 16-7: Presidential Votes *(cont.)*

(a) Using the formula sqrt(.49*.51/n) for the standard deviation of the sample proportion of Clinton voters gives:

n=50	n=100	n=200	n=400	n=500	n=800	n=1000	n=1600	n=2000
.0707	.0500	.0353	.0250	.0224	.0177	.0158	.0125	.0112

(b)



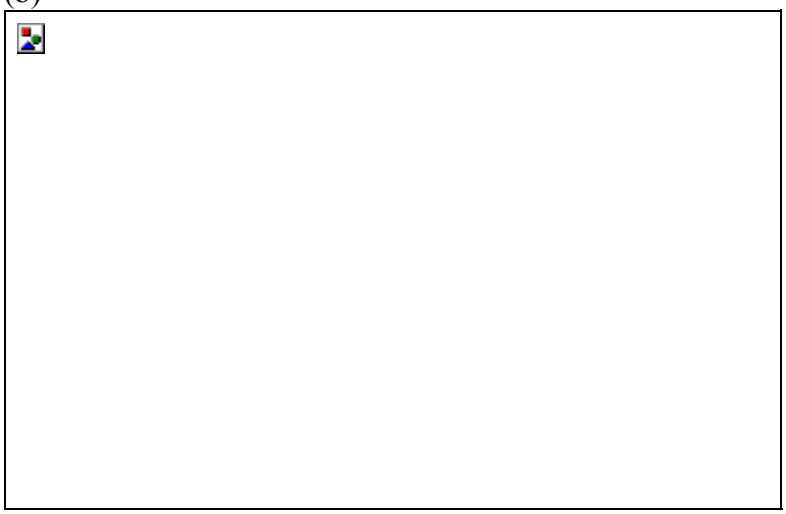
(c) The sample size must increase by a factor of four in order to cut the standard deviation of the sample proportion in half. (For instance, compare n=50 to n=200, or n=100 to n=400, or n=500 to n=2000)

Activity 16-8: Presidential Votes *(cont.)*

(a) Using the formula sqrt(theta*(1-theta)/100) for the standard deviation of the sample proportion of Clinton voters gives:

theta=0	theta=.1	theta=.2	theta=.3	theta=.4	theta=.5	theta=.6	theta=.7	theta=.8	theta=.9	theta=1
0	.03	.04	.0458	.0490	.05	.0490	.0458	.04	.03	0

(b)



(c) Theta = .5 produces the most variability among sample proportions.
(d) Theta = 0 and theta = 1 produce the least variability. If none or all of a population has the characteristic, then none or all (respectively) of the sample will as well, leaving no variability in sample proportions.

Activity 16-9: Cat Households

(a) No, you can not be certain. Even with a smaller sample, your opponent may get lucky.
(b) Yes, with the larger sample you have the better chance of obtaining a sample proportion within a certain distance (such as +/-.05) of the actual population proportion.
(c) When n=200, the standard deviation of the sample proportion is sqrt(.25*.75/200) = .0306. When n=50, this standard deviation is sqrt(.25*.75/50) = .0612. The smaller standard deviation is achieved with the sample size of n=200. It is exactly half as big as when n=50.
d) Results will vary. Expect roughly 90% of the proportions to be within ± .05 of .25.
e) Results will vary but expect median near .25, Q1 near .23 and Q3 near .27.
f) Results will vary. Expect roughly 60% of the proportions to be within ± .05 of .25.
g) Results will vary but expect median near .25, Q1 near .21 and Q3 near .29
h) Results will vary, but here is a sample result:

(i) While they are both centered near .25 and both symmetric, the sample proportions from *n*=50 are more spread out than the sample proportions from *n*=200.

(Activity 16-10: ESP Testing

(a) 15 or more correct identifications are needed so that less than 10% of all guessing subjects do that well. In this simulation, 571 of 10,000 guessing subjects (.0571) scored 15 or higher. Note that 14 is too low, because 1075 of the guessing subjects scored 14 or higher, and this exceeds 10%.
(b) 16 or more correct identifications are needed so that less than 5% of all guessing subjects do that well. In this simulation, 265 of 10,000 guessing subjects (.0265) scored 16 or higher.
(c) 18 or more correct identifications are needed so that less than 1% of all guessing subjects do that well. In this simulation, 49 of 10,000 guessing subjects (.0049) scored

15 or higher. Note that 17 is too low, because 119 of the guessing subjects scored 17or higher, and this exceeds 1%.

Activity 16-11: Calling Heads or Tails

Answers depend on class results.

Activity 16-12: Racquet Spinning *(cont.)*

(a) .46 is a statistic since it is based on the result of a sample of 100 spins.
(b) .50 is a parameter because it refers to the population of all long-run racquet spins.
(c) Answers will vary, but one simulation result is presented here.
(d) The sample proportions of "up" results vary according to a symmetric, bell-shaped distribution. The mean of the sample proportions for this particular simulation is .50065, and the standard deviation is .05190.
(e) The CLT predicts the shape to be normal with mean theta = .5 and standard deviation sqrt(theta*(1-theta)/n) = sqrt(.5*.5/100) = .05. The simulated samples come quite close to the CLT predictions.
(f) In 489 of these 1000 simulated samples (a proportion of .489) is the sample proportion of "up" results either .46 or smaller or .54 or larger.
(g) Since the (approximate simulated) probability in (f) is not at all small, the simulation reveals that the sample result of 46% "up" results is not at all unlikely if the process gives a 50/50 result in the long run.

Activity 16-13: Halloween Practices

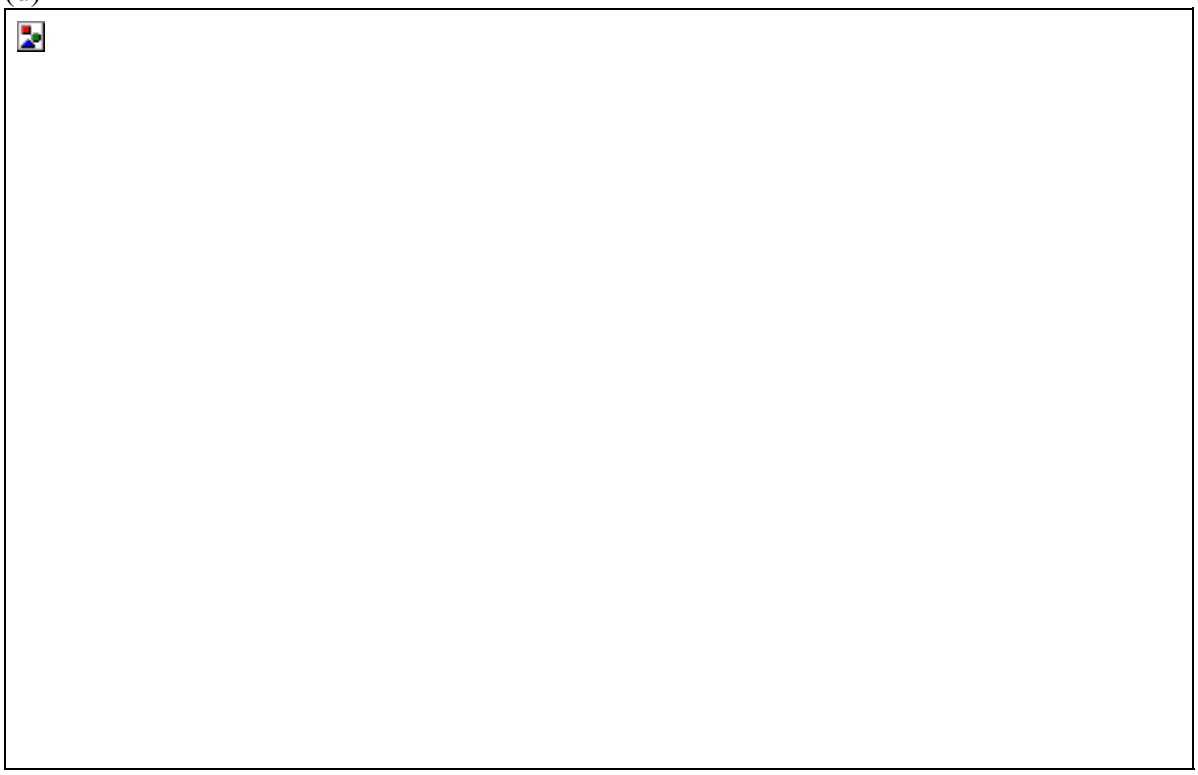
(a) .69 is a statistic, because it is based on a sample of adults interviewed by the Gallup organization.
(b) No. because of sampling variability, the actual population proportion intending to give treats could very well differ from .69 by a bit.
(c) Yes. If theta were .7, the standard deviation of the sample proportions would equal sqrt(.7*.3/1005) = .0145, so 2*.0145 = .029, so sample proportions between .7-.029 = .671 and .7+.029 = .729 would fall within two standard deviations. The observed sample proportion of .69 clearly falls within this interval.
(d) No. If theta were .6, the standard deviation of the sample proportions would equal sqrt(.6*.4/1005) = .0155, so 2*.0155 = .031, so sample proportions between .6-.031 = .569 and .6+.031 = .631 would fall within two standard deviations. The observed sample proportion of .69 clearly does not fall within this interval.
(e) Values of theta for which .69 falls within two standard deviations of theta are .67, .68, .69, .70, and .71. The values .66 and .72 just barely fail to include .69 within two standard deviations.
(f) The sample data suggest that it is most plausible to think that between 67% and 71% of all adult Americans planned to give out treats.

Activity 16-14: Halloween Beliefs

(a)-(d) All of these are possible, but they become less plausible as thge value of theta gets farther from theobserved sample proportion of .22.
(e) Histogram i) is for theta = .23, ii) is for theta = .25, and iii) is for theta = .30. (These can be determined by examining the centers of the simulated sampling distributions.)
(f) 30% is not very plausible, because the simulated sample proportions from theta = .3 (histogram iii) are almost never as low as .22.
(g) 25% is somewhat plausible, because histogram ii) reveals that it is not terribly uncommon for the sample proportion to be as low as .22 when theta = .25.
(h) 23% is quite plausible, because histogram i) reveals that it is quite common to obtain a sample proportion of .22 when theta = .23.

Activity 16-15: Cola Discrimination

(a) A guessing subject would get 1/3 correct in the long run.
(b) You could roll the die repeatedly, with each roll representing a trial. You could let a 1 or 2 represent a correct identification and a 3, 4, 5, or 6 represent an incorrect identification. You would need 30 rolls to represent the thirty trials.
(c) Answers will vary, but one simulation result is presented here.
(d)



(e) The CLT predicts the mean to be equal to theta (1/3, or .3333 in this case) and the standard deviation to be equal to sqrt((1/3)*(2/3)/30) = .0861. For these 1000 simulated sample proportions, the mean is .3368 and the standard deviation is .0864. These are very close to what the CLT predicted.
(f) In these 1000 simulated samples, the guessing subject got 40% or more correct 291 times. This is .291 of the simulated samples.

(g) The simulation reveals that it is not very unusual (prob = about .291) to get 40% or more correct even when one is guessing, so this is not at all convincing evidence that the subject is doing any better than guessing.
(h) Getting 60% or more correct happened for the guessing subject in only 3 of the 1000 simulated samples. Since it is very unusual for a guessing subject to do so well, this result would constitute fairly convincing evidence that the subject was not just guessing.
(i) The distributions are very similar in shape: symmetric and bell-shaped. They are also simialr in variability. The key difference is that one os centered around 1/3 and the other is centered around 2/3. There is some but not much overlap between them.
(j) The subjects gets 60% or more correct in 826 of these 1000 simulated samples, a proportion of .826.