Workshop Statistics: Discovery with Data, Second Edition

Topic 18: Central Limit Theorem

Activity 18-6: Sampling Reese's Pieces *(cont.)*

(a) The sample proportion of orange candies would vary according to a distribution that is approximately normal with mean $\theta = .45$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.45*.55/75) = .0574$.
(c) The z-score for .4 is $z = (.4-.45)/.0574 = -0.87$. Table II reveals that the approximate probability (that less than 40% of the sample will be orange) is .1922.
(d) The z-scores are $(.35-.45)/.0574 = -1.74$ and $(.55-.45)/.0574 = 1.74$. The probability (that between 35% and 55% of the sample is orange) is therefore .9591 - .0409 = .9182.
(e) This probability should be close to that percentage.

Activity 18-7: Sampling Reese's Pieces *(cont.)*

(a) The sample proportion of orange candies would vary according to a distribution that is approximately normal with mean $\theta = .45$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.45*.55/175) = .0376$.
(c) The z-score for .4 is $z = (.4-.45)/.0376 = -1.33$. Table II reveals that the approximate probability (that less than 40% of the sample will be orange) is .0918.
(d) This probability (that less than 40% of the sample will be orange) is smaller with a sample of $n = 175$ than with a sample of $n = 75$ because the larger sample produces a sampling distribution more concentrated around the value .45 and therefore less likely to be as far away as .40 or smaller.
(e) The z-scores are $(.35-.45)/.0376 = -2.66$ and $(.55-.45)/.0376 = 2.66$. The probability (that between 35% and 55% of the sample is orange) is therefore .9961 - .0039 = .9922. This probability is larger with $n = 175$ than with $n = 75$ because the larger sample produces a sampling distribution more concentrated around the value .45 and therefore more likely to be between .35 and .55.

Activity 18-8: ESP Testing *(cont.)*

(a) The CLT reveals that the sample proportion of correct responses will vary according to a distribution that is approximately normal with mean $\theta = .25$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.25*.75/40) = .0685$.
(c) The z-score for .40 is $z = (.40-.25)/.0685 = 2.19$. Table II reveals that the approximate probability (that a guessing subject will correctly identify 40% or more of the cards) is 1-.9857 = .0143.
(d) In the 10,000 simulated samples, the guessing subject got 40% (16) or more correct in 146+70+33+14+0+2 = 265 of them, a percentage of .0265. This is reasonably close to

the approximate probability of .0143.
(e) It would be fairly surprising for a guessing subject to get 40% or more correct on this test. Such a result would happen a little more than 1% of the time in the long run, which is less than 10% but not less than 1%.

Activity 18-9: ESP Testing *(cont.)*

(a) The guessing subject will not always get exactly 25 right. Due to sampling variability, he or she will occasionally get a few more or less than that correct.
(b) The CLT reveals that the sample proportion of correct responses will vary according to a distribution that is approximately normal with mean $\theta = .25$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.25*.75/100) = .0433$.
(c) The z-score for .27 is $z = (.27-.25)/.0433 = 0.46$. Table II reveals that the approximate probability (that a guessing subject will correctly identify 27% or more of the cards) is 1-.6772 = .3228.
(d) This probability is close to 1/3, so it would not be surprising for a subject to get 27% or more correct on this test even when he or she is blindly guessing on each card.
(e) The z-score for .31 is $z = (.31-.25)/.0433 = 1.39$. Table II reveals that the approximate probability (that a guessing subject will correctly identify 31% or more of the cards) is 1-.9177 = .0823. This probability is less than 1/10, so it would be mildly surprising for a subject to get 31% or more correct on this test when he or she is blindly guessing on each card.
(f) The z-score for .35 is $z = (.35-.25)/.0433 = 2.31$. Table II reveals that the approximate probability (that a guessing subject will correctly identify 35% or more of the cards) is 1-.9896 = .0104. This probability is around 1%, so it would be quite surprising for a subject to get 35% or more correct on this test if he or she were blindly guessing on each card.
(g) 33 and 34 each come close to fulfilling this stipulation. The z-score for .33 is 1.85, which produces a probability of .0322. The z-score for .34 is 2.08, which produces a probability of .0188. These are the closest possible probabilities to .025 in this setting.

Activity 18-10: Smoking Rates *(cont.)*

(b) The sampling distribution of the sample proportion of smokers is approximately normal with mean $\theta = .308$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.308*.692/400) = .0231$. The z-score for .333 is $(.333-.308)/.0231 = 1.08$. Table II reveals that the probability that the sample proportion of smokers exceeds .333 is therefore 1-.8599 = .1401. The z-score for .283 is $(.283-.308)/.0231 = -1.08$. Table II reveals that the probability that the sample proportion of smokers is less than .283 is therefore .1401. Thus, the probability that the sample proportion of smokers will be more than .025 away from the population proportion is .1401+.1401 = .2802.
(c) The values of interest here are .308+.05 = .358 and .308-.05 = .258. The respective z-scores turn out to be 2.16 and -2.16. The probability that the sample proportion of smokers will be more than .05 away from the population proportion turns out to be .0154+.0154 = .0308.
(d) Yes, it would be unlikely to get a small sample proportion of smokers if the sample

were chosen randomly from Kentucky residents. Question (c) reveals that the probability of getting 25% or fewer smokers in a random sample of size 400 from Kentucky residents is less than .0154.
(e) Getting a sample proportion of 30% smokers would not be surprising at all in Kentucky, where 30.8% of the population smokes.

Activity 18-11: Christmas Shopping *(cont.)*

(a) With such a small sample, it would not be valid to use the CLT unless the distribution of the amounts themselves were normal. It seems likely that such a variable would not be normal but rather would be skewed to the right, because some people may spend very, very large amounts on Christmas shopping as compared to the majority of people.
(b) Yes, with a sample of size 500, the sample mean would vary according to a distribution that is approximately normal with mean $\mu = 850$ and standard deviation $\sigma/\text{sqrt}(n) = 250/\text{sqrt}(500) = 11.18$.
(c) The z-scores are (868.39-850)/11.18 = 1.64 and (831.61-850)/11.18 = -1.64, so the probability (that the sample mean would fall within $\pm$18.39 of 850) is found from Table II to be .9495-.0505 = .8990, or about 90%.
(d) The z-scores are 1.96 and -1.96, so the probability (that the sample mean would fall within $\pm$21.91 of 850) is found from Table II to be .9750-.0250 = .9500, or 95%.
(e) The z-scores are 2.58 and -2.58, so the probability (that the sample mean would fall within $\pm$28.80 of 850) is found from Table II to be .9951-.0049 = .9902, or about 99%.
(f) The necessary z-score would be 1.28, since the probability of falling between 1.28 and -1.28 in a standard normal distribution is .8997-.1003 = .7994, or about 80%. Thus, 1.28 must equal $k/11.18$, and so $k = 11.18*1.28 = 14.31$.
(g) The z-scores would be 1.64 and -1.64, so the probability would be .8990, just as in (c). The probability of falling within + 18.39 of μ is .8990 regardless of the value of μ.

Activity 18-12: Presidential Votes *(cont.)*

(a) .49 is a parameter because it pertains to the population of all voters in 1996.
(b) The CLT establishes that the sampling distribution of the sample proportion of Clinton voters will vary according to a distribution that is approximately normal with mean $\theta = .49$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.49*.51/500) = .0224$. Thus, the z-score for the value .45 is (.45-.49)/.0224 = -1.79, and Table II reveals that the probability that the sample proportion of Clinton voters would be less than .45 is equal to .0367.
(c) The z-score for the value .5 is (.5-.49)/.0224 = 0.45, and Table II reveals that the probability that the sample proportion of Clinton voters exceeds one-half is equal to 1-.6736 = .3264.
(d) The z-scores are 1.34 and -1.34, so the probability is .9099-.0901 = .8198.
(e) The z-scores are 2.68 and -2.68, so the probability is .9963-.0037 = .9926.
(f) The z-scores are 0 and 17.86, so the probability is 1-.5 = .5.
(g) With a larger sample, the distribution of sample proportions would be more concentrated around the value .49. Thus, the answer to (b) would be smaller, to (c)

would be smaller, to (d) and (e) would be larger, and to (f) would be essentially unchanged.

Activity 18-13: Hospital Births *(cont.)*

(a) The CLT establishes that both hospitals' sampling distributions are approximately normal with mean $\theta = .5$. The standard deviation for hospital A is sqrt(.5*.5/100) = .05, and the standard deviation for hospital B is sqrt(.5*.5/20) = .1118. Thus, hospital B's sampling distribution is much more spread out than hospital A's.
(b) The z-score is (.6-.5)/.05 = 2.00, so the probability of hospital A having 60% or more boy births is 1-.9772 = .0228.
(c) The z-score is (.6-.5)/.1118 = 0.89, so the probability of hospital B having 60% or more boy births is 1-.8133 = .1867.
(d) Hospital B is more likely to have a day with 60% or more boy births.

Activity 18-14: Non-English Speakers

(a) The sampling distribution of the proportion who speak a language other than English at home in California has a distribution that is approximately normal with mean $\theta = .315$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.315*.685/100) = .0465$.
(b) The z-score is (.5-.315)/.0465 = 3.98. Since this value is off the chart of Table II, the probability that more than half of those sampled speak a language other than English at home is less than .0002 (the smallest probability in the table).
(c) The z-score is (.25-.315)/.0465 = -1.40. The probability that fewer than one-quarter of those sampled speak a language other than English at home is therefore .0808.
(d) The z-scores are (.5-.315)/.0465 = 3.98 and (.2-.315)/.0465 = -2.47. The probability that between one-fifth and one-half of those sampled speak a language other than English at home is therefore about .9998-.0068 = .9930.
(e) Ohio's sampling distribution would be approximately normal with mean $\theta = .054$ and standard deviation $\text{sqrt}(\theta(1-\theta)/n) = \text{sqrt}(.054*.946/100) = .0226$. It is less spread out than California's and, more importantly, shifted far to the left (toward lower values) from California's.
(f) The answer to (b) would be even smaller for Ohio, the answer to (c) would be much larger, and the answer to (d) would be much smaller.

Activity 18-15: Solitaire *(cont.)*

(a) In order to satisfy $n\theta \geq 10$ where $\theta = 1/9$, n would have to be at least 90. Note that $n(1-\theta) \geq 10$ is satisfied for smaller values of n, but both conditions must be satisfied for the CLT to hold.
(b) In order to satisfy $n\theta \geq 10$ where $\theta = 1/6$, n would have to be at least 60. Note that $n(1-\theta) \geq 10$ is satisfied for smaller values of n, but both conditions must be satisfied for the CLT to hold.
(c) In order to satisfy $n(1-\theta) \geq 10$ where $\theta = .8$, n would have to be at least 50. Note that $n\theta \geq 10$ is satisfied for smaller values of n, but both conditions must be satisfied for the

CLT to hold.