Review Problems
Chapter 23 — Accuracy of Averages

Main point: The formulas for a percentage and an average are similar.
A percentage is an average for a box model composed of zeros and ones,
an average can be based on any arbitrary box model (incomes of families, SAT scores, age ...)
For an average based on N random draws,
   EV of average = average of box
   SE of average  =  (sqrt N) * SD of box / N  =  SD of box / (sqrt N)

Simulation exercise: average of 400 draws from arbitrary box:

   box <- c(1, 2, 2, 5)
   mean(box) = 2.5
   SD(box) = 1.5
   means <- c()  # Placeholder for accumulating means from simulation
   for (i in 1:10000) means[i] <- mean(draw(400 box))
   mean(means) = 2.4994  (EV would be 2.5)
   SD(means) = 0.0748    (SE would be 1.5 / (sqrt 200) = 1.5 / 20 = 0.0750

* Problem 1. A box has mean 100 and SD 20; an average of 400 draws with replacement from the box will have
   EV of the average = 100
   SE of the average = 20 / (sqrt 400) = 20 / 20 = 1
   a. It is virtually certain that the average will be in the range 100 +/- 20 SE units
   b. There is a 68 percent probability that another draw of 400 will give an average of 100 +/- 1 SE unit.

* Problem 2. We have a box with an unknown average and SD, but with 500 draws we have
   a sample average of 71.3 and and SD of 2.3.
   We use the bootstrapping method to construct a box identical to our sample, and consider the
   likely variation in averaging another 500 draws:
   EV of the average of another 500 draws = 71.3 (equal to mean of box)
   SE of the average of another 500 draws = 2.3 / (sqrt 500) = 2.3 / 22.3607 = .1029
   a. Our estimate of the population average is EXACTLY 71.3; this is the EXACT average of the
      box we construct to model the population. It is not likely to be the true population average.
   b. The 68 percent confidence interval for the average is in fact 71.3 +/- 0.1029;
      this is based on the Central Limit Theorem which tells us that whatever the distribution of the
      population, the sums, percentages and averages tend to be distributed normally.
      Hence if we sample repeatedly from the box, we expect 68 percent of the averages to
      fall within this range.
   c. We have no information at all on the distribution of tickets in the box. If we were sure that the
      distribution were normal, 68 percent of the tickets would be in the range 71.3 +/- 2.3
      (NOT +/- 0.1)
* Problem 3. Real estate survey on commuting distances.
  
  Random sample of 1000 households to find commuting distance of heads of household to work.

Average of the sample = 8.7 miles with SD of 9.0 miles. Note that the distribution is certainly skewed -- you can't commute less than zero miles. The box model will be a bootstrapped version of the sample, with mean of 8.7 miles and SD 9.0 miles.

If we take another sample of 10,000 households from our "bootstrapped" box, we would have:
  
  EV of average = 8.7 miles
  SE of average = 9.0 / (sqrt 1000) = 0.2846

So we estimate the average commute distance as 8.7 miles, give or take 0.2846 miles.

A 90 percent confidence interval for our average is 8.7 +/- 1.645 * 0.2846 miles = 8.7 +/- 0.4682 miles
  or from 8.2318 to 9.1682.

For a 95 percent confidence interval, the exact multiplier would be 1.96 rather than 1.645.

The convention of looking for a 95 percent CI comes from the greater convenience of multiplying by 2 to come up with an approximately 95 percent CI.

Problem 4. Cluster sampling.

The real estate survey also looked at all other members of the household. Doing this, simply because it is more convenient to do so, is often financially desirable (see chapter 22, which indicates that the Bureau of Labor Statistics and the Census do so to compute unemployment). But it makes it difficult to be sure of what the standard error is. Households may contain only one person, or they may contain several teenagers and aged grandparents -- the sample of households may differ from a simple random sample of the population.

If we were assured that the differences were minor, we could apply the formulas:

Average for the city will have an EV of 7.7 with SE 10.2 / (sqrt 2500) = 0.2040, so a VERY approximate 95 percent CI would be 7.7 +/- 0.4. (The text authors are probably rigid on this point, and would prefer the answer "No, you can't compute the CI." I'd accept "Be very, very careful of trusting the CI").

* Problem 5. Percentage of car commuters.

Sample: 721 of 1000 heads of households commute by car (randomly sampled).

Box: 721 ones, 279 zeros. Mean of box = 0.721 or 72.1 percent;
  
  SD of box = sqrt (0.721 * 0.279) = 0.4485

Hence EV of percentage = 0.721 or 72.1 percent.

  SE of percentage = 0.4485 / (sqrt 1000) = 0.0142 or 1.42 percent

and 95 % CI for the percentage of car commuters = 72.1 +/- 2.84 = 69.26 to 74.94


Sample data: sample size = 1000; average score = 307; SD of scores = 30.

IF sample is random, likely size of the chance error ( = 1 SE) is about 30 / (sqrt 1000) = 0.9487

so a 95 percent CI would be about 28 to 32.

IF sample is not strictly random, we cannot be sure that any CI we construct is accurate -- warnings of problem 4 apply.

Problem 7. Seriously stressed people.

The sample has serious problems with self-selection bias, since those who take the questionnaire may well differ from the general population. Those who feel more stressed may welcome the opportunity to complain, or the very, very stressed may feel they already have more than enough to do -- and we cannot be sure that the biases will cancel out. There is no point in trusting the survey results AT ALL, let alone calculating a confidence interval. If the sample were random, the SE would be .0093 or about .9 percent -- but it ISN'T.
**Review Problems -- Chapter 23 -- Accuracy of Averages (p. 3)**

* Problem 8. Survey of Colleges.

Four hundred colleges are randomly chosen, with average enrollment = 3,700; SD = 6,500.
We use the initial sample for our box model, and on its basis calculate the EV and SE of an average of size 400.

EV of the average = 3,700
SE of the average = 6500 / (sqrt 400) = 325

a. "A 68 percent CI for the average of ALL colleges will be 3700 +/- 325" is very tempting to say, but is not accurate -- the average enrollment of all colleges is what it is, and does not have a confidence interval around it.
b. "If we take repeated samples of size 400, we expect 68 percent of the averages will be within the range 3700 +/- 25" is a true statement.
c. "About 68 percent of the schools in the sample have a student enrollment of 3700 +/- 6500"
   Confuses SD and SE; wrongly assumes normality (there are many more small schools than large schools); and implies that there are many schools with negative enrollments.
   The chance a school's enrollment is less than zero is given by first finding the Z-score:
   
   \[
   \frac{(0 - 3700)}{6500} = -0.5692 \text{ and } (\text{normal-cdf} -0.5692) = 0.2846,
   \]
   
   implies that 28.5 percent of US colleges actually have negative enrollments!
d. "About 68 percent of schools have enrollments of 3700 +/- 325". Using the SE instead of the SD to describe schools is also wrong -- the SE applies to averages, not to actual schools.
   Statement (c) would have been correct IF the distribution of schools were normal, and IF negative enrollments were possible -- this will be right only by accident.
e. "Since the distribution of schools is not normal, we can't make any statement about confidence intervals" Wrong again --
   BECAUSE of the Central Limit Theorem, we can always make inferences about averages, which are expected to follow the normal distribution.

Problem 9. Faculty publications.

Looking at the schools chosen above, simple random sample of 2,500 faculty members in the 400 colleges previously selected were chosen, with average number of research papers = 1.7 and SD = 2.3.
If we mechanically apply the formula, the SE will be 2.3 / (sqrt 2500) = .0460, permitting a 95 percent CI of about 1.7 +/- .09.
But we have a cluster sample and hence cannot be sure that the SE has been accurately calculated. See pages 202-3 for discussion.

* Problem 10. Persons per household, based on sample of size 625.

Sample results = bootstrapped box = Average family size = 2.30 and SD of family size = 1.75
(a) The SE for an average from size 625 is (sqrt 625) * 1.75 / 625 = 0.07, as claimed
(b) The average household size in the sample we actually took IS 2.30; no CI is possible.
(c) The 95 percent CI for the average household size in the city is 2.30 +/- 0.14, as claimed.
(d) ZERO percent of the households in the city contain between 2.16 and 2.44 persons.
   Many may contain 2, and many may contain 3 -- but the CI applies to AVERAGES, not DATA.
   (e) Household size is NOT normal. If it were, 2.5 percent of the families would have fewer than 2.30 - 2 (1.75) = 2.30 - 3.50 = MINUS 1.2 members.

Problem 11. The scale on the X-axis is wrong: Mean of box is 3, and SD of box is 1.414, so SE of average of 25 numbers is 1.414 / (sqrt 625) = 0.2828, so each unit on the X-axis covers 3.5 SE units -- from 2 to 4 should cover 99.953 percent of the data. More of the data is outside this range -- visually it seems to be about 5 percent.
Use EcLS to take a look at (normal-area -3.5 3.5) and (normal-area -2 2), and compare to the histogram.

Problem 12. Not a CI because we don't have a random sample of either students or final exams.