Homework - Week 7

Solution

Questions marked with "(OS3: X.X)" are from the textbook with "X.X" as the exercise number. The answers to the odd questions (odd by book numbering that is) will be in the back of the book.

1. (OS3: 4.1) For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

   a. In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

      Numeric variable, parameter is a mean.

   b. In a survey, one hundred college students are asked: "What percentage of the time you spend on the Internet is part of your course work?"

      Numeric variable, parameter is a mean.

   c. In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

      Categorical (binary) variable, parameter is a proportion.

   d. In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

      Numeric variable, parameter is a mean.

   e. In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

      Categorical (binary) variable, parameter is a proportion.

2. (OS3: 4.3) A college counselor is interested in estimating how many credits a student typically enrolls in each semester. The counselor decides to randomly sample 100 students by using the registrar's database of students. The histogram in the book shows the distribution of the number of credits taken by these students. Sample statistics for this distribution are also provided.

Min	8
Q1	13
Median	14
Mean	13.65
SD	1.91
Q3	15
Max	18

   a. What is the point estimate for the average number of credits taken per semester by students at this college? What about the median?

      Point estimate = sample mean = 13.65.

   b. What is the point estimate for the standard deviation of the number of credits taken per semester by students at this college? What about the IQR?

      Point estimate = sample SD = 1.91.

   c. Is a load of 16 credits unusually high for this college? What about 18 credits? Explain your reasoning. *Hint:* Observations farther than two standard deviations from the mean are usually considered to be unusual.

2 SD above the mean = 13.65 + 1.91 * 2 = 17.47. Thus, 16 credits is not unusual, but 18 credits is unusually high.

d. The college counselor takes another random sample of 100 students and this time finds a sample mean of 14.02 units. Should she be surprised that this sample statistic is slightly different than the one from the original sample? Explain your reasoning.

No. We would expect sample statistics to vary between random samples.

e. The sample means given above are point estimates for the mean number of credits taken by all students at that college. What measures do we use to quantify the variability of this estimate (Hint: recall that $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample.

Standard deviation of sample means is $\frac{\sigma}{\sqrt{n}}$. Sample SD is used to estimate σ. Thus, SD of sample means = 1.91/sqrt(100) = 0.191.

3. (OS3: 4.4) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram in the book shows the sample distribution of heights in centimeters.

Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

a. What is the point estimate for the average height of active individuals? What about the median?

Point estimate = sample mean = 171.1.

b. What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

Point estimate = sample SD = 9.4.

c. Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

```
z.180 <- (180-171.1)/9.4
z.180
```

```
## [1] 0.9468085
```

```
z.155 <- (155-171.1)/9.4
z.155
```

```
## [1] -1.712766
```

Both z-scores are within less than 2 and greater than -2. Thus, neither height should be considered unusual.

d. The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

No. We would expect sample statistics to vary between random samples.

e. The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity

using the data from the original sample under the condition that the data are a simple random sample.

Standard deviation of sample means is $\frac{\sigma}{\sqrt{n}}$. Sample SD is used to estimate σ. Thus, SD of sample means = 9.4/sqrt(507) = 0.42.

4. (OS3: 4.5) The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means.

   a. What is this distribution called?

      The sampling sistribution.

   b. Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

      Because two SD below the mean is negative, an impossible value for counts of eggs, the distribution is likely to be right-skewed.

   c. Calculate the variability of this distribution and state the appropriate term used to refer to this value.

      Standard deviation of sample means, or standard error, is $\frac{\sigma}{\sqrt{n}}$. Standard error = 18.2/sqrt(45) = 2.71.

   d. Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

      Standard error = 18.2/sqrt(10) = 5.76.

5. (OS3: 4.7) In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions". However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting. Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

   95% confidence interval $= p \pm z_{0.05} \times SE = 0.45 \pm 1.96 \times 0.012 =$ **(0.42648, 0.47352)**

   We are 95% confident that the true proportion of adults that report living with one or more cronic conditions is between 42.6% and 47.3%.

6. (OS3: 4.9) In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions", and the standard error for this estimate is 1.2%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

   a. We can say with certainty that the confidence interval from Question 5 (4.7) contains the true percentage of U.S. adults who suffer from a chronic illness.

      False. We are only 95% confident that the interval contains the true proportion.

   b. If we repeated this study 1,000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of U.S. adults who suffer from chronic illnesses.

      True. This is another valid interpretation of the confidence interval.

   c. The poll provides statistically significant evidence (at the $\alpha = 0.05$ level) that the percentage of U.S. adults who suffer from chronic illnesses is below 50%.

True. Since the values in the confidence interval are all below 50%, this is true (although the significance is greater than 0.05).

d. Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

False. Standard error has nothing to do with uncertainity of answers.

7. (OS3: 4.12) The 2010 General Social Survey asked the question: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

   a. Interpret this interval in context of the data.

   We are 95% confident that the true mean number of days that US residents felt their mental health was not good, out of the past 30 days, is between 3.4 and 4.24.

   b. What does "95% confident" mean? Explain in the context of the application.

   If we repeated this process 100 times, we would expect the resulting confidence intervals to contain the true population mean about 95 times.

   c. Suppose the researchers think a 99% confidence level would be more appropriate for this interval. Will this new interval be smaller or larger than the 95% confidence interval?

   Because we want to be more confident our interval contains the true value, our confidence interval would be larger.

   d. If a new survey were to be done with 500 Americans, would the standard error of the estimate be larger, smaller, or about the same. Assume the standard deviation has remained constant since 2010.

   With a smaller sample size, the standard error will be larger (we are dividing standard deviation by a smaller number).

8. (OS3: 4.16) The National Survey of Family Growth conducted by the Centers for Disease Control gathers information on family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health. One of the variables collected on this survey is the age at first marriage. The histogram in the book shows the distribution of ages at first marriage of 5,534 randomly sampled women between 2006 and 2010. The average age at first marriage among these women is 23.44 with a standard deviation of 4.72.

   Estimate the average age at first marriage of women using a 95% confidence interval, and interpret this interval in context. Discuss any relevant assumptions.

```
ci <- 23.44 + c(-1, 1) * 1.96 * 4.72 / sqrt(5534)
ci
```

```
## [1] 23.31564 23.56436
```

We are 95% confident that the true mean age of first marriage in this population of women is between 23.32 and 23.56 years.