Mat1372 HW2

5.1 Identify the parameter, Part I. 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.

(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?"

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

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

(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.

Sol: (a) mean
(b) mean
(c) proportion (Yes or No question)
(d) mean
(e) proportion (Yes No question)

5.2 Identify the parameter, Part II. For each of the following situations, state whether the parameter of interest is a mean or a proportion.

(a) A poll shows that 64% of Americans personally worry a great deal about federal spending and the budget deficit.

(b) A survey reports that local TV news has shown a 17% increase in revenue within a two year period while newspaper revenues decreased by 6.4% during this time period.

(c) In a survey, high school and college students are asked whether or not they use geolocation services on their smart phones.

(d) In a survey, smart phone users are asked whether or not they use a web-based taxi service.

(e) In a survey, smart phone users are asked how many times they used a web-based taxi service over the last year.

Sol: (a) proportion (worry or not)
(b) mean
(c) proportion (whether or not)
(d) proportion (whether or not)
(e) mean

5.3 Quality control. As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(a) What population is under consideration in the data set?

(b) What parameter is being estimated?

(c) What is the point estimate for the parameter?

(d) What is the name of the statistic we use to measure the uncertainty of the point estimate?

(e) Compute the value from part (d) for this context.

(f) The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?

(g) Suppose the true population value was found to be 10%. If we use this proportion to recompute the value in part (e) using $p = 0.1$ instead of $\hat{p}$, does the resulting value change much?

Sol: (a) population: all chips at this factory during this week

(b) defective rate

(c) The defective rate is $\hat{p} = \frac{27}{212} = 0.127$

(d) We quantify this uncertainty using the standard error. (SE)

(e) $SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127 \cdot (1-0.127)}{212}} = 0.023$

(f) For historical rate 10%, we have

$\frac{0.1 - 0.127}{0.023} = -1.1739$ which is around one SE away from $\hat{p}$. So it is not unusual.

(g) If $p = 0.1$. Then $SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1 \cdot 0.9}{212}} = 0.021$

and the value is not that different

5.4 Unexpected expense. In a random sample 765 adults in the United States, 322 say they could not cover a $400 unexpected expense without borrowing money or going into debt.

(a) What population is under consideration in the data set?

(b) What parameter is being estimated?

(c) What is the point estimate for the parameter?

(d) What is the name of the statistic we use to measure the uncertainty of the point estimate?

(e) Compute the value from part (d) for this context.

(f) A cable news pundit thinks the value is actually 50%. Should she be surprised by the data?

(g) Suppose the true population value was found to be 40%. If we use this proportion to recompute the value in part (e) using $p = 0.4$ instead of $\hat{p}$, does the resulting value change much?

Sol: (a) Population: All adults in the state

(b) The proportion that an adult can't cover a $400 unexpected expense

(c) $\hat{p} = \frac{322}{765} = 0.421$

(d) The standard Error

(e) $n = 765$. $\hat{p} = 0.421$. Then

$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.421 \cdot (1-0.421)}{765}} = 0.0179$$

but p is unknown

(f) For 50%, we have

$\frac{0.5 - 0.421}{0.0179} = 4.413$ which is more than 4 SE away from $\hat{p}$. It means that 50% is uncusual.

(g) If $p = 0.4$, we have

$SE_p = \sqrt{\frac{0.4(1-0.4)}{212}} = 0.0177$ which is not much different from the SE in (e).

5.5 Repeated water samples. A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

(a) What is this distribution called?

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

(c) If the proportions are distributed around 8%, what is the variability of the distribution?

(d) What is the formal name of the value you computed in (c)?

(e) Suppose the researchers' budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?

Sol (a) sampling distribution

(b) Sample size $n = 800$, $\hat{p} = 5\% \sim 30\%$, then we have

$40 < n\hat{p} < 240 \Rightarrow n\hat{p} > 10$

$560 < n(1-\hat{p}) < 760 \Rightarrow n(1-\hat{p}) > 10$

} It passes success-failure condition.

If the population proportion p is in this range (5% ~ 30%), then the distribution should be symmetric.

(c) We can use the standard error to see how the distribution spreads when $p = 0.08$:

$$SE = \sqrt{\frac{0.08(1-0.08)}{800}} = 0.0096$$

(d) Standard Error

(e) The distribution will tend to be more variable when we have fewer observations per sample.

5.7 Chronic illness, Part I. In 2013, the Pew Research Foundation reported that "45% of U.S. adults report that they live with one or more chronic conditions".[11] 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.

Sol: Sample proportion: $\hat{p} = 0.45$
Standard error $SE_{\hat{p}} = 0.012$ which is normal distributed.

95% confidence interval: $\hat{p} \pm 1.96 \cdot SE$

$\rightarrow 0.45 \pm 1.96 \times 0.45 \rightarrow (0.426, 0.474)$

It means that we are 95% confident that proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%

5.9 Chronic illness, Part II. 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 Exercise 5.7 contains the true percentage of U.S. adults who suffer from a chronic illness.

(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.

(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%.

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

Sol (a) False, Confidence intervals provide a range of plausible values, and the truth sometimes is missed (which is 5% of the time)

(b) True. This is the exact definition of 95% confidence interval

(c) True, since 50% is not in (42.6%, 47.4%), then if this is in a hypothesis test with $H_0: p = 50\%$, H_0 will be rejected.

(d) False, the SE means the uncertainty in the overall estimate from natural fluctuations due to randomness.

5.8 Twitter users and news, Part I. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter.[12]. The standard error for this estimate was 2.4%, and a normal distribution may be used to model the sample proportion. Construct a 99% confidence interval for the fraction of U.S. adult Twitter users who get some news on Twitter, and interpret the confidence interval in context.

Sol: sample proportion $\hat{p} = 52\% = 0.52$, standard error $SE_{\hat{p}} = 2.4\% = 0.024$ } and it is normal distributed.

We have a 99% confidence interval as $\hat{p} \pm 2.58 \times SE_{\hat{p}}$

$\Rightarrow 0.52 \pm 2.58 \times 0.024 \Rightarrow (0.458, 0.582)$ and

it means that we have 99% confident that 45.8% to 58.2% of US adult Twitter users get some news from Twitter.

5.10 Twitter users and news, Part II. A poll conducted in 2013 found that 52% of U.S. adult Twitter users get at least some news on Twitter, and the standard error for this estimate was 2.4%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(a) The data provide statistically significant evidence that more than half of U.S. adult Twitter users get some news through Twitter. Use a significance level of $\alpha = 0.01$. (This part uses concepts from Section 5.3 and will be corrected in a future edition.)

(b) Since the standard error is 2.4%, we can conclude that 97.6% of all U.S. adult Twitter users were included in the study.

(c) If we want to reduce the standard error of the estimate, we should collect less data.

(d) If we construct a 90% confidence interval for the percentage of U.S. adults Twitter users who get some news through Twitter, this confidence interval will be wider than a corresponding 99% confidence interval.

Sol. (a) False, If we build a hypothesis with $H_0: p = \frac{1}{2}$ (exactly 50% of US adult Twitter uses get news on Twitter), then it will not be rejected since 50% is included in the 99% confidence interval. Thus, we can't say it is more than 50% (which is $H_A: p \neq \frac{1}{2}$)

(b) False, the SE means the uncertainty in the overall estimate from natural fluctuations due to randomness, not about the percentage of sample in the study

(c) False, to reduce the SE, we need more data

(d) False, for 90% confidence interval, we have a smaller $z^* = 1.96$ and, thus, the interval is smaller than the 99% one.

5.11 Waiting at an ER, Part I. A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning.

(a) We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes.

(b) We are 95% confident that the average waiting time of all patients at this hospital's emergency room is between 128 and 147 minutes.

(c) 95% of random samples have a sample mean between 128 and 147 minutes.

(d) A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.

(e) The margin of error is 9.5 and the sample mean is 137.5.

(f) In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size. (Hint: the margin of error for a mean scales in the same way with sample size as the margin of error for a proportion.)

Sol: This should be put in ch7. ... I'll come back to this in ch7.