Problem Set 5

Danielle Sockin

2023-10-11

Question 1

Confidence interval: point estimate $\pm$ margin of error(z or t statistic multiplied by the standard error)

a) The level of confidence reflects the degree of precision associated with our methods. It will affect the size (width) of a confidence interval because it will impact the z- or t-statistic we choose, and therefore will impact the margin of error. For instance, a confidence level of 90% has the z-statistic 1.645, a confidence level of 95% has a z-statistic 1.96, and a 99% confidence interval has a z-statistic of 2.58. The larger the desired level of confidence/precision, the wider the confidence interval will be as a mathematical function of multiplying by increasingly higher numbers to calculate the margin of error that gets added to and subtracted from the point estimate.

b) The sample standard deviation (s) will affect the size (width) of a the confidence interval as part of the margin of error. The larger the sample standard deviation, the larger the confidence interval. This makes sense mathematically because s is the numerator in our standard error calculation ($\frac{s}{\sqrt{n}}$); as s increases, holding else constant, so too will the standard error and therefore the margin of error, widening the confidence interval.

c) The sample size (n) will affect the size (width) of a confidence interval as part of the margin of error. The larger the sample size, the smaller the confidence interval. This makes sense numerically because n is the denominator in our standard error calculation ($\frac{s}{\sqrt{n}}$); as n increases, holding else constant, the denominator will increase, lowering the standard error and therefore the margin of error, narrowing the size of the confidence interval.

Question 2

a) Using the t-table handout, if n=31, the critical value of t you would use to make a 95% confidence interval for a sample mean is 2.042. I got this number by calculating the degrees of freedom as n-1 (31-1=30) and by using a right-tail probability of .025 (half of the remaining 5%).

b) With a t-statistic of 2.799 and a degrees of freedom equal to 26, the area in the upper-tail of the t-distribution is between 0.001 and 0.005.

c) If there are 60 degrees of freedom and the t-statistic is -2.42. the area in the lower tail of the t-distribution is between 0.005 and 0.010.

Question 3

Survey sample size 650; mean years of education 8.2; standard deviation (s) 5.5

a) 95% confidence interval around the sample mean: $\bar{y} \pm$ t $\frac{s}{\sqrt{n}}$ = 8.2 $\pm$ 1.96$\frac{5.5}{\sqrt{650}}$ = 8.2 $\pm$ 0.4228 = (7.777, 8.623). Interpretation: The calculated 95% confidence interval is 7.777 to 8.623. Across repeated samples, there is a 95% chance that a confidence interval constructed with these procedures will include the true population mean of years of education for people in the unspecified country.

b) 90% confidence interval around the sample mean: $\bar{y} \pm$ t $\frac{s}{\sqrt{n}}$ = 8.2 $\pm$ 1.645$\frac{5.5}{\sqrt{650}}$ = 8.2 $\pm$ 0.3549 = (7.845, 8.555). Interpretation: The calculated 90% confidence interval is 7.846 to 8.553. Across repeated samples, there is a 90% chance that a confidence interval constructed with these procedures will include the true population mean of years of education for people in the unspecified country.

c) A decade earlier, the country found a mean of 7.9 years of education. It is plausible that 7.9 years is still the mean of the population. Relating this to parts a and b, 7.9 falls within both the 90 and 95 percent confidence intervals. There is no guaranteeing that our confidence intervals contain the true population mean (the 90 and 95 percent are indications of our precision) but because they both contain 7.9, there is not enough evidence to suggest that 7.9 is no longer the population mean.

Question 4

Random sample of 110 cars; proportion of 0.15 failed to meet emissions standards

a) 90% confidence interval associated with estimated proportion of 0.15: $\hat{\pi} \pm$ z $\sqrt{\frac{(\hat{\pi})(1-\hat{\pi})}{n}}$= 0.15 $\pm$ 1.645 $\sqrt{\frac{0.15*0.85}{110}}$ = 0.15 $\pm$ 1.645(0.0340) = 0.15 $\pm$ 0.0559 = (0.094, 0.206)

b) The calculated confidence interval is from 0.094 to 0.206. Across repeated samples, there is a 90% chance that a confidence interval constructed with these measures will include the true population proportion of cars that fail to meet emissions standards.

c) α (alpha) = 0.05, null hypothesis is that the true proportion of the cars that fail emission standards is 0.06

- considering assumptions: Our sample is random as given in the question. We don't have enough observations in each category - 0.06(110) = 6.6 - to apply the CLT so going forward, we need to assume that the sample is normally distributed to proceed with the significance test even though or n does not satisfy necessary requirements. We will use a z-test because we are working with proportions.
- stating hypotheses: The null hypothesis is that true proportion of cars that fail emissions standards is 0.06. The alternative hypothesis is that the true proportion of cars that fail emissions standards is not 0.06.
  H_0: $\pi = 0.06$; H_A: $\pi \neq 0.06$
- identifying critical value of the test statistic: $\alpha = 0.05$ at 95% confidence interval so the critical value of z will be 1.96 (with area of 0.025 in each tail)
- calculating value of the test statistic for our sample proportion:z = $\frac{\hat{\pi}-\pi_0}{\sigma\pi_0}$ = $\frac{0.15-0.06}{0.0226}$ = 3.982
  $\sigma\pi 0 = \sqrt{\frac{(\pi_0)(1-\pi_0)}{n}} = \sqrt{\frac{(.06)(1-.06)}{110}} = \sqrt{\frac{0.0564}{110}} = 0.0226$
- make a decision on whether to accept or reject the null hypothesis: Because our z test statistic is equal to 3.982, which is greater than our critical value of z which is 1.96, at the 0.05 level of significance we reject the null hypothesis that the true proportion of cars failing to meet emission standards is 0.06.
- calculate p-value: If we round to our z test statistic from 3.982 to 4.0, the p-value is (2)(0.0000317) = 0.0000634.
  If we don't round then we should say the p-value is between (2)(0.000233) and (2)(0.0000317) so between 0.0000634 and 0.000466 (won't change the main result of rejecting the null hypothesis based on an alpha level of 0.05).

Question 5

anes2020subset dataset - variable BAplus asked respondents whether they have a college degree: yes or no

```
library(questionr)
freq(table(anes2020$BAplus), cum=T, valid=F)
```

```
##                    n    %  %cum
## Does not have BA 4502 55.2  55.2
## Has BA           3647 44.8 100.0
```

```
prop.test(3647, 8149, conf.level = .90, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  3647 out of 8149, null probability 0.5
## X-squared = 89.707, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.4384982 0.4566158
## sample estimates:
##         p
## 0.4475396
```

a) The proportion of respondents who responded yes was 0.448

b) 95% confidence interval associated with estimated proportion of 0.448: $\hat{p} \pm \text{z} \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} = 0.448 \pm 1.96 \sqrt{\frac{(0.448)(1-0.448)}{(3647+4502)}} = 0.448 \pm 1.96(0.0055) = 0.448 \pm 0.0108 = (0.4372, 0.4588)$

c) Our calculated confidence interval is 0.437 to 0.4588. Across repeated samples, there is a 95% chance that a confidence interval constructed with these procedures will include the true population proportion of people that have a college degree.

d) Our sample proportion reflects one sample of the population. Across repeated random samples, confidence intervals following certain procedures will have an X% chance (like 95% or 99%) of containing the true population proportion of people in the population that have a college degree. Consequently, there will always be a chance that a particular confidence interval does not contain the true population proportion and it is possible that our sample falls into that bucket.

   Additionally, our interval might not contain the true proportion of people in the population that have a college degree if our sample proportion is influenced by response bias. If people in our sample are inclined to give untrustworthy/inaccurate answers to the question of if they have a college degree (because they are embarassed or because not having a college degree can be looked down upon), then our sample proportion would be calculated on biased responses. A confidence interval based upon a biased sample proportion $\hat{p}$ might not contain the true population proportion.

e) The 90% confidence interval is (0.4385, 0.4566). Across repeated samples, there is a 95% chance that a confidence interval constructed with these procedures will include the true population proportion of people that have a college degree.

Question 6

Treating scales from anes2020subset as interval-level variables; 50 means neutral, goes from 0 to 100

```
mean(anes2020$UnionsTherm, na.rm=TRUE)
```

```
## [1] 58.33511
```

```
sd(anes2020$UnionsTherm, na.rm=TRUE)
```

```
## [1] 24.0318
```

```
length(anes2020$UnionsTherm[!is.na(anes2020$UnionsTherm)])
```

```
## [1] 7326
```

```
t.test(anes2020$UnionsTherm, conf.level = .99)
```

```
##
##  One Sample t-test
##
## data:  anes2020$UnionsTherm
## t = 207.77, df = 7325, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  57.61170 59.05852
## sample estimates:
## mean of x
##  58.33511
```

a) The mean thermometer score for police (UnionsTherm) is 58.335, the sample size is 7326 after removing NAs (overall sample is 8280 observations), and the standard deviation is 24.032.

b) 95% confidence interval around the sample mean: $\bar{y} \pm$ t $\frac{s}{\sqrt{n}} = 58.335 \pm 1.96\frac{24.032}{\sqrt{7326}} = 58.335 \pm 1.96(0.2808) = (57.785, 58.885)$

c) The 99% confidence interval for UnionTerm is (57.612, 59.059). Across repeated samples, there is a 95% chance that a confidence interval constructed with these procedures will include the true population mean thermometer score for police (UnionsTherm).