Exam 02

STA 198 Spring 2020 (Jiang)

Wednesday, July 8, 2020

Question 2

2.1

The total width of a confidence interval for the mean is given by 2 times the margin of error. Given that the underlying distribution is known to be normal with known variance, we can use the standard normal distribution as our reference distribution. Thus, for a 95% interval and substituting in known values, we have

$$
\begin{aligned}
L &= 2Z^{\star}_{0.975}\frac{\sigma}{\sqrt{n}} \\
n &= \left(\frac{2Z^{\star}_{0.975}\sigma}{L}\right)^2 \\
&\approx \left(\frac{2 \times 1.96 \times \sqrt{6}}{1.5}\right)^2 \\
&\approx 40.98.
\end{aligned}
$$

Rounding up, we have that 41 samples are required to sure that the length of the interval is at most 1.5.

2.2

Once again plugging in, we have

$$
\begin{aligned}
L &= 2Z^{\star}\sigma\sqrt{n} \\
Z^{\star} &= \frac{L\sqrt{n}}{2\sigma} \\
&\approx \frac{1.5 \times \sqrt{50}}{2 \times \sqrt{6}} \\
&= 2.165.
\end{aligned}
$$

```
pnorm(2.165)
```

```
## [1] 0.9848062
```

The critical value of 2.165 corresponds to the 98.48th percentile of the normal distribution. Working backwards, we have $1 - 2 \times (1 - 0.9848) = 0.9696$. Thus, any confidence interval with confidence level approximately 96.96% or below will have total length of at most 1.5, given a sample size of 50 and a variance of 6.

2.3

The length of this interval is 0.52. Once again plugging in, we have

$$
\begin{aligned}
L &= 2Z^{\star}_{0.975}\frac{\sigma}{\sqrt{n}} \\
n &= \left(\frac{2Z^{\star}_{0.975}\sigma}{L}\right)^2 \\
&\approx \left(\frac{2 \times 1.96 \times \sqrt{6}}{0.52}\right)^2 \\
&\approx 341.
\end{aligned}
$$

341 observations were used.

2.4

We know that if we were to independently take new samples of the same size and from the same underlying population and then construct 95% confidence intervals in the same manner as our original case, then we would expect 95% of those intervals to truly contain the population mean.

2.5

Lower than 0.05.

2.6

In Q2.5, Research Team B was essentially testing the null hypothesis that $\mu = 4$ vs. the alternative hypothesis that $\mu \neq 4$. Given that their 95% confidence interval did not contain the null hypothesized value and that we are testing a mean, they would have obtained a p-value lower than 0.05, which would have been rejected.

The researchers thus would have concluded that there is sufficient evidence that the mean LPW in this valley was not equal to 4 mm. In this case, the researchers' conclusion would have been consistent with the truth that the mean LPW is in fact 3.65 mm. No error was made.

Note: although the researchers would have estimated the mean LPW to be 3.54 which is not 3.65, with their observed data they would not have had sufficient evidence to reject the hypothetical null hypothesis that $\mu = 3.65$ (the actual true value).

Question 3

Questions 3.1, 3.2, 3.4, 3.8, and 3.9 are TRUE. Questions 3.3, 3.5, 3.6, 3.7, and 3.10 are FALSE.

Question 4

4.1

First, let's create a table examining the number of first draws over 15 ppb.

```
library(tidyverse)
flint <- read_csv("flint.csv")

flint %>%
```

```r
  mutate(over15 = ifelse(lead > 15, "over", "under")) %>%
  filter(draw == "first") %>%
  count(over15)
```

```
## # A tibble: 2 x 2
##   over15     n
##   <chr>  <int>
## 1 over      45
## 2 under    226
```

It appears 45 out of 271 draws were over 15 ppb. Now, create the confidence interval:

```r
prop.test(x = 45, n = 271, conf.level = 0.95)
```

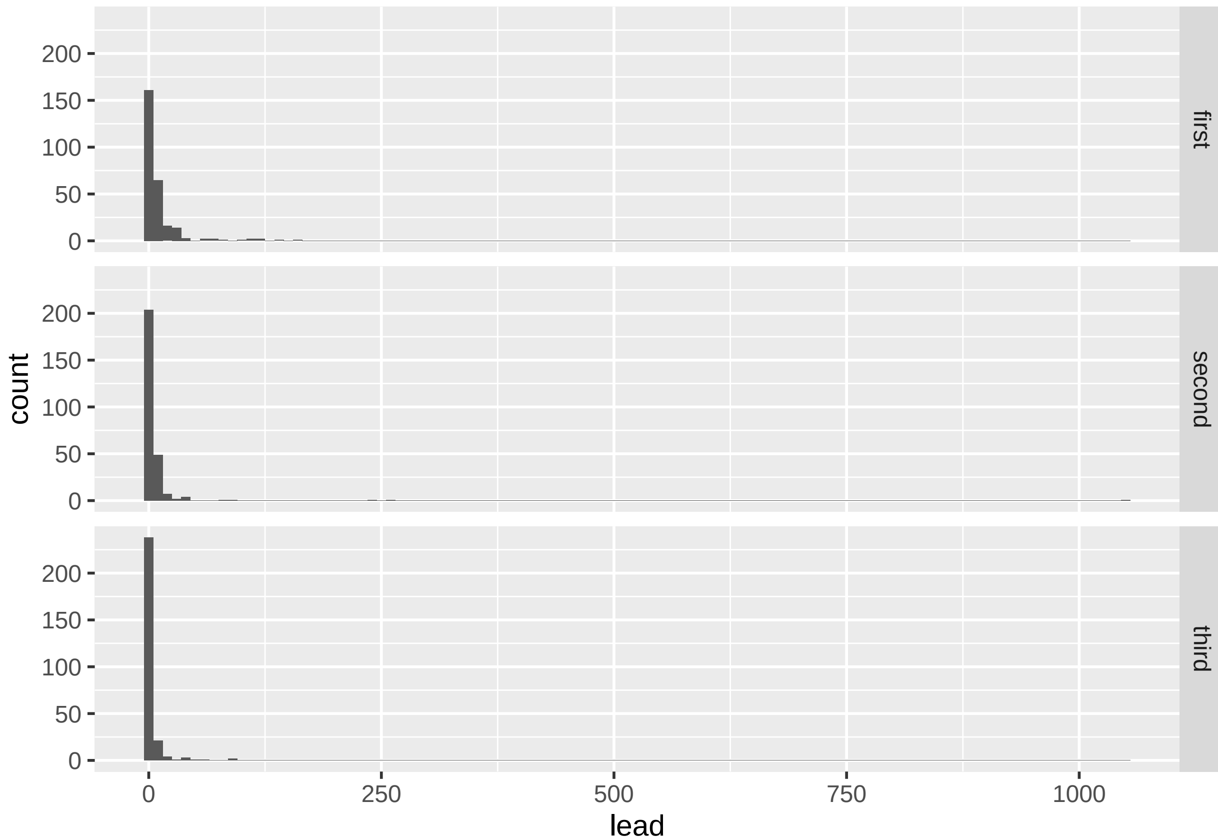
```
##
##  1-sample proportions test with continuity correction
##
## data:  45 out of 271, null probability 0.5
## X-squared = 119.56, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1248515 0.2169709
## sample estimates:
##         p
## 0.1660517
```

The 95% confidence interval is approximately (0.125, 0.217).

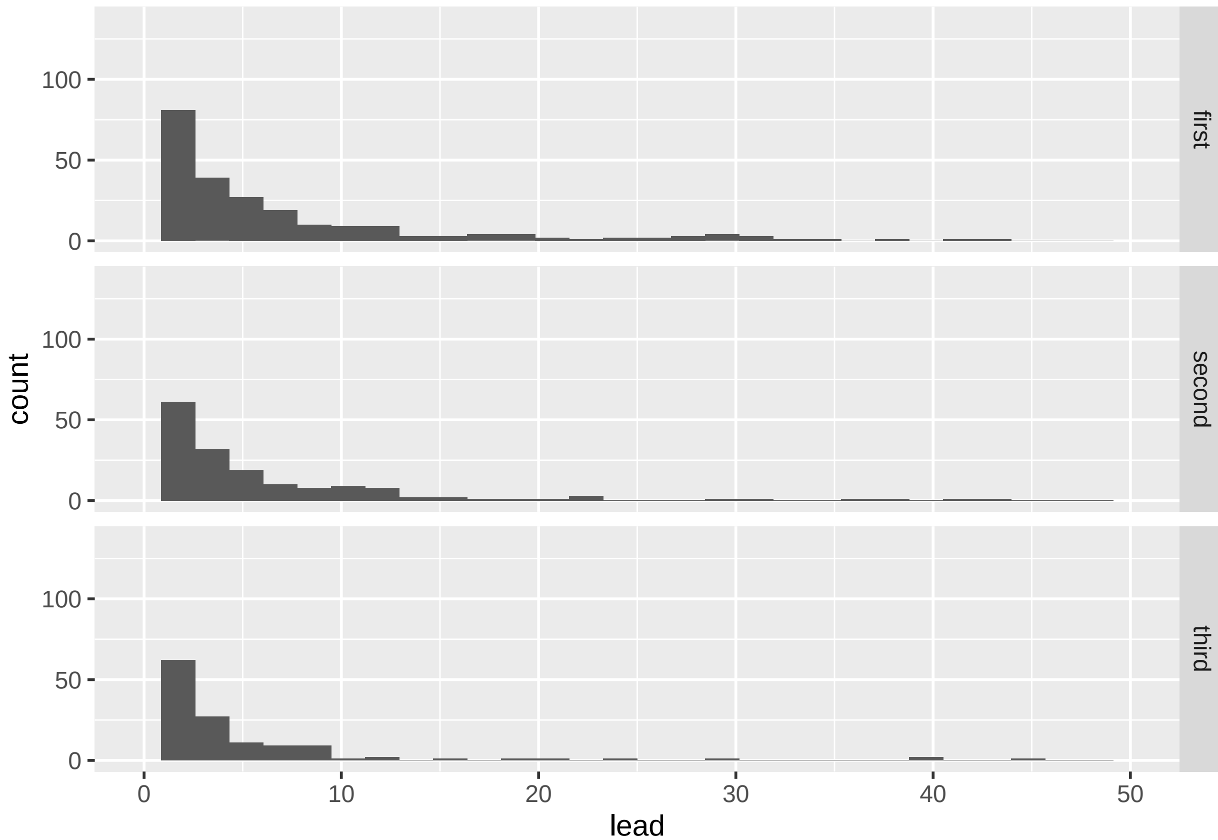
4.2

We aim to evaluate a continuous variable by three categories. Thus, either an ANOVA or Kruskal-Wallis test would be appropriate, depending on the situation. Let's evaluate the assumptions for ANOVA. We first plot the distribution of lead levels by which draw the sample was taken from and summarize the variance of the lead values by each group. Because of the extreme outliers, we will also provide a "zoomed-in" plot to examine lower values.

```r
ggplot(data = flint, aes(x = lead)) +
  geom_histogram(binwidth = 10) +
  facet_grid(draw~.)
```




```
ggplot(data = flint, aes(x = lead)) +
  geom_histogram() +
  facet_grid(draw~.) +
  xlim(c(0, 50))
```




```r
flint %>%
  group_by(draw) %>%
  summarize(var_lead = var(lead))
```

```
## # A tibble: 3 x 2
##   draw   var_lead
##   <chr>     <dbl>
## 1 first      465.
## 2 second    4560.
## 3 third      111.
```

Although it may be reasonable that independence is satisfied (one resident's lead levels are unlikely to affect anothers; they're probably both affected by which city water main their houses are on), it is clear that normality of observations within each group is not satisfied (heavy right skew) and that the groups do not have the same variance. Thus, we select the Kruskal-Wallis test.

```r
kruskal.test(lead ~ draw, data = flint)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  lead by draw
## Kruskal-Wallis chi-squared = 120.76, df = 2, p-value < 2.2e-16
```

The null hypothesis for our test is that all three draw groups have the same median lead levels; the alternative is that at least one group has a different median lead level from the others. Our test statistic was 120.76, which has a χ^2_2 distribution under the null hypothesis, corresponding to a p-value < 0.001. At the $\alpha = 0.05$ significance level, we reject the null hypothesis; there is sufficient evidence to suggest that at least one group has a different median lead level than the others.

The appropriate step-down test is a rank sum test testing the null hypothesis that median lead levels are the same in each draw category being compared vs. the alternative that median lead levels are not the same in these two categories.

```
flint %>%
  filter(draw != "first") %>%
  wilcox.test(lead ~ draw, data = .)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  lead by draw
## W = 43918, p-value = 7.872e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
flint %>%
  filter(draw != "second") %>%
  wilcox.test(lead ~ draw, data = .)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  lead by draw
## W = 56490, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
flint %>%
  filter(draw != "third") %>%
  wilcox.test(lead ~ draw, data = .)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  lead by draw
## W = 49328, p-value = 4.657e-12
## alternative hypothesis: true location shift is not equal to 0
```

All three pairwise comparisons were < 0.001 and thus significant at the a Bonferroni-corrected α level of $0.05/3$. Thus, we reject each hypothesis; there is sufficient evidence to suggest that the median lead levels are not the same between the first and second draws, the first and third draws, and the second and third draws.

4.3

We perform a one-sample test of proportion, testing whether the proportion of first draw samples having over 15 ppb of lead is greater than 10%. Our null hypothesis is that the true proportion of such samples is ≤ 0.1 and our alternative hypothesis is that the true proportion is > 0.1 (note that a two-sided hypothesis of the proportion being equal to 0.1 under the null and not equal to 0.1 under the alternative is also acceptable).

From before, we found that 45 of the 271 values were greater than 15 ppb. Thus, the assumptions for our test of proportion are satisfied, as there are both enough "over" and "under" values in our dataset. Carrying out this test, we have:

```
prop.test(x = 45, n = 271, p = 0.1, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  45 out of 271, null probability 0.1
```

```
## X-squared = 12.413, df = 1, p-value = 0.0002131
## alternative hypothesis: true p is greater than 0.1
## 95 percent confidence interval:
##  0.1305442 1.0000000
## sample estimates:
##         p
## 0.1660517
```

Our test statistic was 12.4, which has a χ^2_1 distribution under the null hypothesis, corresponding to a p-value of approximately 0.0002. Thus, at the $\alpha = 0.05$ level, we reject the null hypothesis. There is sufficient evidence to suggest that the proportion of first draw samples having lead above 15 ppb is greater than 0.1.

Note: a test against the standard normal distribution would also have been acceptable. As well, a two-sided test would have been fine as well.

4.4

In examining the lead levels in water samples, we find that the median amount of lead in the sample was associated with how long the tap has been running. In particular, it appears that the longer the tap has been running, the lower the lead level. Although the MDEQ did not find evidence that lead levels were dangerously elevated, these samples were based on letting the tap run for five minutes, which we would expect to have lower lead levels than would reflect in reality. Additionally, we found that there was sufficient evidence to suggest that the water quality in Flint exceeded the EPA guidance for actionable lead levels (greater than 10% of first draw samples having lead levels greater than 15 ppb). Thus, there appears to be evidence that Flint residents are being exposed to dangerously high levels of lead in their drinking water and recommend further investigation and/or remediation.

Question 5

5.1

A chi-square test would be appropriate for this situation, as we see that there are two categorical variables (educational attainment vs. death due to cancer) and all cell counts have >5 observations in the contingency table. The null hypothesis is that there is no association between educational attaiment vs. death due ot cancer, and the alternative hypothesis is that such an association does exist. Under the null hypothesis, our test statistic follows a χ^2_3 distribution. Given the p-value of 0.50, which is greater than our pre-specified α level of 0.05, we fail to reject the null hypothesis. We have insufficient evidence to suggest that there is an association between educational level and death due to cancer.

5.2

We see that seven tests are being performed. Assuming that these tests are independent, our family-wise type I error rate would be $1 - (0.95)^7$, or approximately 0.302.