Homework 3

Matthew Bentz

3/24/2022

Part I: Knowledge-Based Questions

1. You have been hired by customer quality assurance firm to investigate claims of quality control fraud. Colgate claims that a tube of toothpaste has 16 oz of toothpaste in it. Let this be your null hypothesis $H_0 : \mu = 16$. To test this hypothesis, you conducted a sample of 40 tubes of toothpaste. and found a 95% confidence interval of $[14.3, 16.7]$. Do you reject / fail to reject / cannot make a decision for the null hypothesis at an $\alpha = .05$ level of significance? What about at $\alpha = .1$? $\alpha = .01$?

At $\alpha = .05$ significance, we can fail to reject the null hypothesis because we know that $\mu = 16$ falls in the interval given.
At $\alpha = .1$ significance, we cannot make a decision for the null hypothesis because the confidence interval becomes smaller, but we do not know if it includes $\mu = 16$.
At $\alpha = .01$ significance, we can fail to reject the null hypothesis because the confidence interval gets larger, but we already know our null hypothesis is within the $\alpha = .05$ significance.

2. What happens to the error of a confidence interval as the sample size increases if the level of confidence and standard deviation are not changed?

The error of a confidence interval decreases as the sample size increases.

Part II: Application Questions

3. Heavy metal pollution of various ecosystems is a series ecological threat. One study showed that in a sample of size 56 Mugil liza fish species, the sample average of zinc concentration was 9.15 $\mu g/g$ with a sample standard deviation of 1.27 $\mu g/g$. Calculate the 95% confidence interval for the population mean concentration of zinc.

n = 56
$\bar{x} = 9.15$
S = 1.27
$\alpha = .05$

```r
# Find the t-values associated with the 95% confidence interval.
t <- qt(.025, 55)

# Find the low/high values on the confidence interval.
low_interval <- 9.15 - (2 * 1.27 / sqrt(56))
high_interval <- 9.15 + (2 * 1.27 / sqrt(56))
```

The 95% CI for the population mean is [8.81, 9.49].

4. Quality assurance testing at the Chips Ahoy! factory was testing the claim that there are 1300 chocolate chips per bag. They wanted to see if there were actually less than 1300 chips as advertised. A random sample of 30 bags yielded a mean of 1261.6 chips per bag and a sample standard deviation of 117.6 chips. Construct a hypothesis test using a .05 level of significance ($\alpha = .05$) for if there are on average less than 1300 chips per bag. Use the critical value approach to make a decision.

H_0: $\mu = 1300$
H_a: $\mu < 1300$ (left-tailed)
$n = 30$
$\bar{x} = 1261.6$
$S = 117.6$
$\alpha = .05$

$t_0 = \frac{\bar{x}-\mu_0}{\frac{S}{\sqrt{n}}} = -1.788$
$t_\alpha = qt(\frac{\alpha}{2}, n-1) = -2.045$

```
tzero <- (1261.6 - 1300) / (117.6 / sqrt(30))
talpha <- qt(0.05, 29)
```

We conducted a one-sample hypothesis test for the population mean of chips per bag. We got a test statistic of $t_0 = -1.788$ with 29 degrees of freedom. This value is greater than the absolute value of the critical value $t_\alpha = -1.699$, so we reject H_0 that the population mean is on average less than 1300 chips per bag at a level of significance $\alpha = .05$.

5. Quality assurance testing at the Chips Ahoy! factory was testing the claim that there are 1300 chocolate chips per bag. They wanted to see if there were actually less than 1300 chips as advertised. A random sample of 30 bags yielded a mean of 1261.6 chips per bag and a sample standard deviation of 117.6 chips. Construct a hypothesis test using a .05 level of significance ($\alpha = .05$) for if there are on average less than 1300 chips per bag. Use the p-value approach to make a decision.

H_0: $\mu = 1300$
H_a: $\mu < 1300$ (left-tailed)
$n = 30$
$\bar{x} = 1261.6$
$S = 117.6$
$\alpha = .05$

$t_0 = \frac{\bar{x}-\mu_0}{\frac{S}{\sqrt{n}}} = -1.788$

```
tzero <- (1261.6 - 1300) / (117.6 / sqrt(30))
p <- pt(tzero, 29)
```

We conducted a one-sample hypothesis test for the population mean of chips per bag. We got a test statistic of $t_0 = -1.788$ with 29 degrees of freedom. The p-value for -1.788 is .042. This value is less than α (.05), so we reject H_0 that the population mean is on average less than 1300 chips per bag at a level of significance $\alpha = .05$.

Part III: Bootstrap Confidence Intervals

The following questions in this Part will all involve the data gathered by Erich Brandt's study involving tipping at a restaurant. 20 patrons were randomly sampled and found that their tip percentages where:

```
tip_data = c(22.7, 13.5, 16.3, 18.9, 13.6,
             20.2, 27.6, 16.0, 16.1, 40.1,
             13.6, 20.2, 29.9, 18.2, 22.8,
             16.0, 19.0, 19.2, 14.0, 15.7)
```

6. Find the sample average for tip_data.

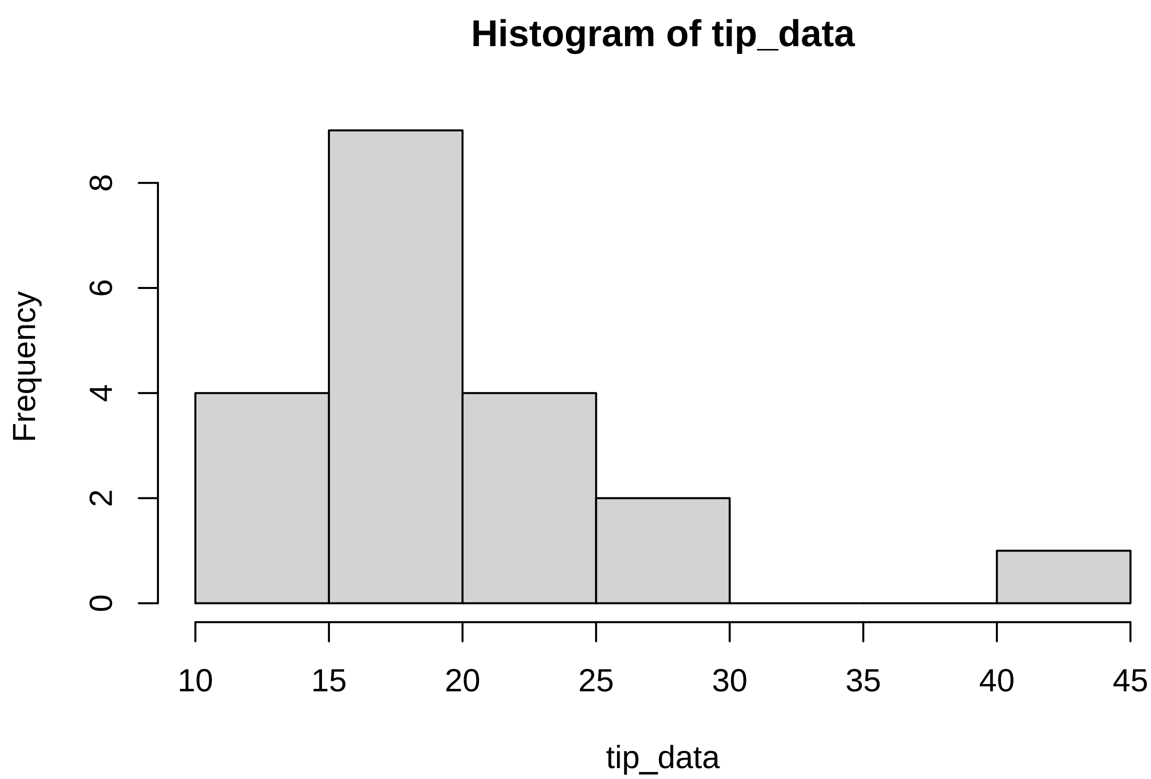
```
mean(tip_data)
```

```
## [1] 19.68
```

The sample average is 19.68.

7. Is the data normally distributed as required by the assumptions? If it is not normal, comment on the shape, spread, and skewness of the distribution. Can we still assume normality of the sampling distribution? Why or why not?

```
hist(tip_data)
```



The data is not normally distributed. The histogram shows that the tip data does have resemblance to a bell-shaped graph, but it is not symmetrical (due to the outlier of 40%). The spread of the graph is from 10-45%, but the mean is close to 20%. The distribution is also skewed to the right because of this outlier.

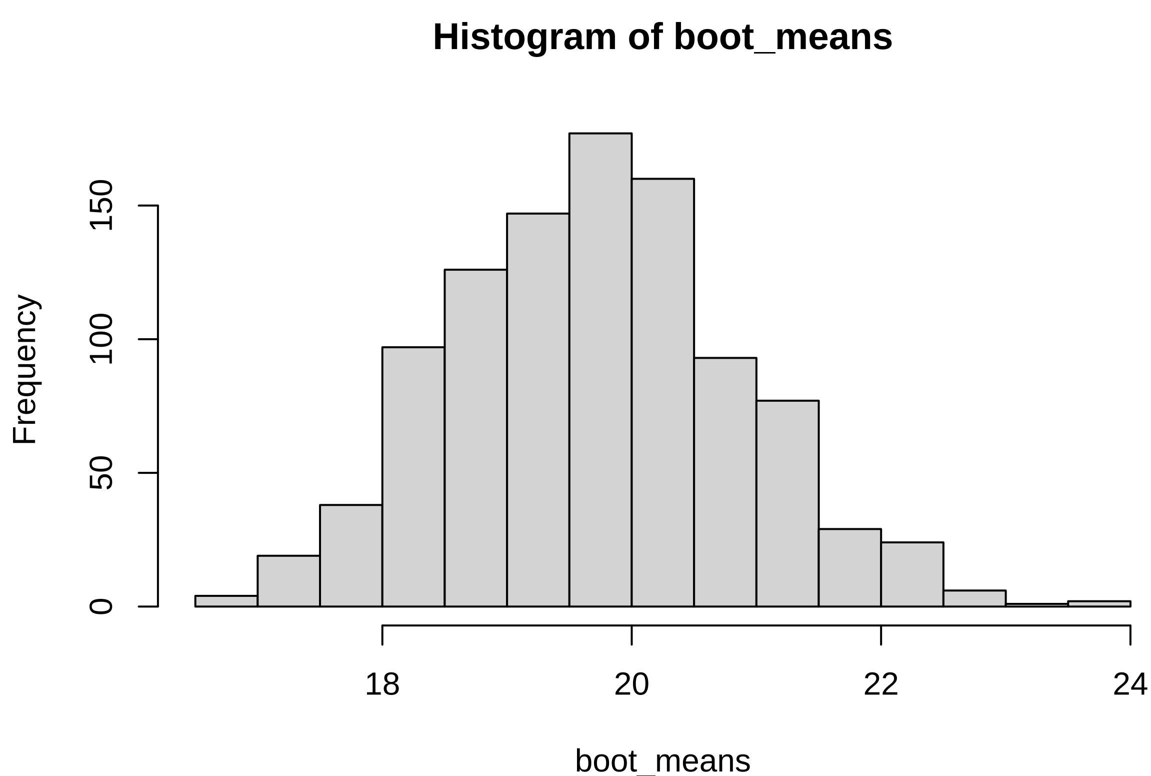
Since our assumptions are violated, we are unable to make a confidence interval using the methods we learned in lecture. However, statisticians proved that you can resample with replacement from the data enough times and "pull the data up by it's bootstraps" to be able to do analysis. The following code takes 1000 sample means from samples of size 30 that are drawn from our original dataset with replacement. Run the following code before you do the next problems.

```
boot_data = NA
boot_means = NA

for(i in 1:1000){
  boot_data = sample(tip_data, 30, replace = T)
  boot_means[i] = mean(boot_data)
}
```

8. Look at the histogram for the boot_means variable. Compare this histogram to the histogram for the tip_data. Where is this histogram centered? Is it close or far away from the mean of tip_data?

```
hist(boot_means)
```



The histogram is centered around 20. This is close to the mean of tip_data (19.68).

9. Construct a 98% confidence interval using $\overline{x_{boot}} \pm z^*_{\alpha/2} s_{boot}$ where $\overline{x_{boot}}$ and s_{boot} are the sample mean and standard deviation for the boot_means and $z^*_{\alpha/2}$ is found using $qnorm()$.

```r
x <- mean(boot_means)
s <- sd(boot_means)

# Auto-generating the interval.
qnorm(.01, mean = mean(boot_means), sd = sd(boot_means))
```

```
## [1] 17.01203
```

```r
qnorm(.99, mean = mean(boot_means), sd = sd(boot_means))
```

```
## [1] 22.40906
```

```r
# Finding the Z values.
z_neg <- qnorm(.01)
z_pos <- qnorm(.99)

# Manually finding the intervals.
low_interval2 <- x + z_neg * s
high_interval2 <- x + z_pos * s
```

The 98% confidence interval, $\overline{x_{boot}} \pm z^*_{\alpha/2} s_{boot}$, is [16.99, 22.44].

The interval you just found is called a bootstrap confidence interval and it's one of the ways statisticians are able to make confidence intervals for true population means even when the normality assumption is not satisfied.