R Lab 7

Bryce Tondryk

April 28

R LAB

Hypothesis tests for one proportion

The `prop.test()` function is used to perform a one-sample proportion test. The function is called as follows:

```
prop.test(x, n, p, conf.level = 0.95, alternative = c("less", "two.sided", "greater"),
          correct = FALSE)
```

Note that R uses `x` to denote the number of successes rather than `y`. The code `conf.level = 0.95` shows that the default confidence level is 0.95. Note that the default for the alternative hypothesis is `"two.sided"`. By default, R uses Yates' continuity correction in the `prop.test()` function, and this is useful if the expected number of successes or failures is fewer than 5. This is beyond the theory of this class, so we set `correct = FALSE` to compute the uncorrected z-test of a proportion, as we have done in class.

One proportion test with $H_1 : p < p_0$

Let's consider the following example. Suppose that nationally, 90% of home buyers would select a home with energy efficient features, even paying a premium of 2-3%, to reduce their utility costs long term. The president of an engineering firm that installs these energy efficient features wants to know if less than 90% of buyers in her local area feel this way (because if so, she might want to launch an education campaign to help her business). She hires you to randomly sample 360 potential buyers in her area, and you find that 312 of them would indeed select a home with energy efficient features.

You want to test your sample against the null hypothesis that the true population proportion (for the engineering firm's area) is 0.90, with the alternative that the proportion is less than 0.90. We test the null hypothesis $H_0 : p = 0.90$ against the alternative hypothesis $H_1 : p < 0.90$ using the `prop.test()` function in R as follows:

```
prop.test(x = 312, n = 360, p = 0.9, alternative = "less", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  312 out of 360, null probability 0.9
## X-squared = 4.4444, df = 1, p-value = 0.01751
## alternative hypothesis: true p is less than 0.9
## 95 percent confidence interval:
##  0.000000 0.893418
## sample estimates:
##         p
## 0.8666667
```

The input for `prop.test()` is the number of successes, followed by the number of trials, and then the value p_0 from the null hypothesis $H_0 : p = p_0$. In the `prop.test()` function, the default values for the confidence level and the alternative hypothesis are `conf.level = 0.95` and `alternative = "two.sided"`. If you wish to use a different confidence level, then you must include this argument in the `prop.test()` function as well; we will see this in another example.

The output for `prop.test()` given above includes

1. the value of Pearson's chi-squared test statistic.
2. p-value.
3. 95% confidence interval.
4. the point estimate for p, or $\widehat{p}$.

To make a conclusion for the hypothesis test, you must use the p-value, as the observed value of the test statistic is not computed here. Note that Pearson's chi-squared test statistic is beyond the scope of this class.

EXERCISE 1

State your conclusion for the hypothesis test above with a significance level of $\alpha = 0.05$, based on the computations given by the `prop.test()` function. Interpret your results in terms of the problem. That is, do you advise the president of the engineering firm to begin the aforementioned education campaign?

Answer: Since the p(value) $= 0.01751 < 0.05 = \alpha$, we reject H_0. This means that less than 90% of home buyers would buy a house at premium with energy efficient features even if that would reduce cost in the long term- we estimated this with 95% confidence of course. This means that the president should launch the aforementioned campaign.

One proportion test with $H_1 : p > p_0$

Because of tourism in Michigan, it was proposed that public schools in the state begin after Labor Day. To determine whether support for this change was greater than 65%, a public poll was taken. Let p equal the proportion of Michigan adults who favor a post-Labor Day start. We shall test $H_0 : p = 0.65$ against $H_1 : p > 0.65$ with a significance level of $\alpha = 0.025$, given that 414 out of a sample of 600 favor a post-Labor Day start.

```
prop.test(414, 600, p = 0.65, conf.level = 0.975, alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  414 out of 600, null probability 0.65
## X-squared = 4.2198, df = 1, p-value = 0.01998
## alternative hypothesis: true p is greater than 0.65
## 97.5 percent confidence interval:
##  0.6518829 1.0000000
## sample estimates:
##    p
## 0.69
```

EXERCISE 2

State your conclusion for the hypothesis test above regarding the start date for public schools, based on the computations given by the `prop.test()` function. Interpret your results in terms of the problem.

Answer: Since the P(value) = .01998 < .025 = α, we also reject H_0. This signifies that with 97.5% confidence, the percentage of support for public schools starting after labor day was greater than 65%. —

Hypothesis tests for two proportions

The `prop.test()` function is also used to perform a two-sample proportion test. The function is called as follows:

```
prop.test(x = c(x1, x2), n = c(n1, n2), conf.level = 0.95,
          alternative = c("less", "two.sided", "greater"), correct = FALSE)
```

As before, `conf.level = 0.95` shows that the default confidence level is 0.95, and the default for the alternative hypothesis is `"two-sided"`. In order to utilize `prop.test` for a two proportion hypothesis test, we simply set `x` and `n` equal to vectors representing the two values for each population.

Two proportion hypothesis test with $H_1 : p \neq p_0$

A machine shop that manufactures toggle levers has both a day and a night shift. A toggle lever is defective if a standard nut cannot be screwed onto the threads. Let p_1 and p_2 be the proportion of defective levers among those manufactured by the day and night shifts, respectively. We shall test the null hypothesis $H_0 : p_1 = p_2$ against a two-sided alternative hypothesis with a significance level of $\alpha = 0.05$. This will be based on two random samples, each of 1000 levers taken from the production of the respective shifts. In the samples, $y_1 = 37$ and $y_2 = 53$ defectives were observed for the day and night shifts, respectively. We run the hypothesis test using the `prop.test()` function.

```
prop.test(x = c(37, 53), n = c(1000, 1000), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(37, 53) out of c(1000, 1000)
## X-squared = 2.9785, df = 1, p-value = 0.08438
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.034157139  0.002157139
## sample estimates:
## prop 1 prop 2
##  0.037  0.053
```

EXERCISE 3

State your conclusion for the hypothesis test above regarding defective levers from each shift, based on the computations given by the `prop.test()` function. Interpret your results in terms of the problem.

Answer: Since p-value = 0.08438 > 0.05 = α, we fail to reject H_0. That is we cannot say with 95% confidence that our alternative is true, that is the defects of the day and night shift could be equal.

Hypothesis tests for means

In class, we learned about different hypothesis tests for one mean, depending on the scenario:

- Population variance known (z-test)
- Population variance unknown with sample size $n > 30$ (z-test)

- Population variance unknown with sample size $n \leq 30$ (t-test, $n-1$ df)
- X and Y are dependent samples. For $W = X - Y$, we test the null hypothesis that $\mu_W = 0$ versus an alternative. (paired t-test, $n-1$ d.f.)

In applications, the reality is that the population variance is rarely known. In addition, as the number of degrees of freedom for a t-distribution gets larger and larger, the t-distribution approaches a normal distribution. This can be seen in the t table in the textbook (page 505) where the $r \to \infty$ row matches up with a $N(0, 1)$ distribution. For these reasons, using a t-test in R will be sufficient for any hypothesis tests we desire to run.

First, we will use the `midwest` data frame from the `ggplot2` package for one and two-sample independent t-test examples. Use the console to view the `midwest` data frame and read the help file before continuing.

`t.test` function

The `t.test()` function can be used to perform both one and two sample t-tests on vectors of data. The function contains a variety of arguments and is called as follows:

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Here `x` is a numeric vector of data values and `y` is an *optional* numeric vector of data values. If `y` is excluded, the function performs a one-sample t-test on the data contained in `x`, if it is included it performs a two-sample t-test using both `x` and `y`.

The `mu` argument provides a number indicating the true value of the mean (or difference in means if you are performing a two sample test) under the null hypothesis. That is, for a one-sample test, this is really μ_0 in the null hypothesis $H_0 : \mu = \mu_0$. By default, the test performs a two-sided t-test; however, you can change the alternative argument to "greater" or "less" depending on which alternative hypothesis you would like to test. For example, the code below performs a one-sample t-test on the data contained in `x` with $H_0 : \mu = 25$ and $H_1 : \mu < 25$.

```
t.test(x, alternative = "less", mu = 25)
```

The `paired` argument will indicate whether or not you want a paired t-test. The default is set to `FALSE` but can be set to `TRUE` to perform a paired t-test.

The `var.equal` argument indicates whether or not to assume equal variances when performing a two-sample t-test. The default `FALSE` value assumes unequal variance and applies Welch's approximation to the degrees of freedom; however, you can set this to TRUE to pool the variance. Although we did not cover Welch's approximation in class due to tedious computations by hand, this approximation should be used when the sample variances are seemingly much different, particularly when the samples sizes are also relatively small different than each other.

Finally, the `conf.level` argument determines the confidence level of the reported confidence interval for μ in the one-sample case and $\mu_1 - \mu_2$ in the two-sample case.

One-sample t-test

The one-sample t-test compares a sample's mean with a known value when the variance of the population is unknown. Suppose we want to assess the percent of college educated adults in the Midwest and compare it to a certain value. For example, let's assume the nation-wide average of college educated adults is 35% (Bachelor's degree or higher) and we want to see if the Midwest mean is significantly different than the national average; in particular we want to test if the Midwest average is less than the national average. We use a significance level of $\alpha = 0.005$ to compute the confidence interval.

```
t.test(midwest$percollege, mu = 35, alternative = "less", conf.level = 0.995)
```

```
##
```

```
##  One Sample t-test
##
## data:  midwest$percollege
## t = -55.842, df = 436, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 35
## 99.5 percent confidence interval:
##      -Inf 19.04771
## sample estimates:
## mean of x
##  18.27274
```

In this code, we begin with `midwest$percollege` because the variable (column) `percollege` of the `midwest` data frame is the vector of data we wish to run the t-test on. Note that the `t.test()` function returns the observed value of the test statistic, unlike the `prop.test()`.

EXERCISE 4

At a significance level of $\alpha = 0.005$, would you reject $H_0 : \mu = 35$ and accept $H_1 : \mu < 35$, or would you fail to reject H_0? Interpret your answer in terms of the problem.

Answer: Since the p-value = 2.2e-16 $< .005 = \alpha$, we reject H_0. This means that with 99.5% confidence we believe that the percentage of Midwest adults who have a bachelor's degree is less than 35%, hence we accept H1.

Two-sample t-test

Now let's say we want to compare the differences between the average percent of college educated adults in Ohio versus Michigan. Here, we want to perform a two-sample t-test. First, we filter the data frame to only include the Ohio and Michigan data, and then only the `state` and `college` variables. We also begin exploring the data by viewing the five-number summary of the data for each state.

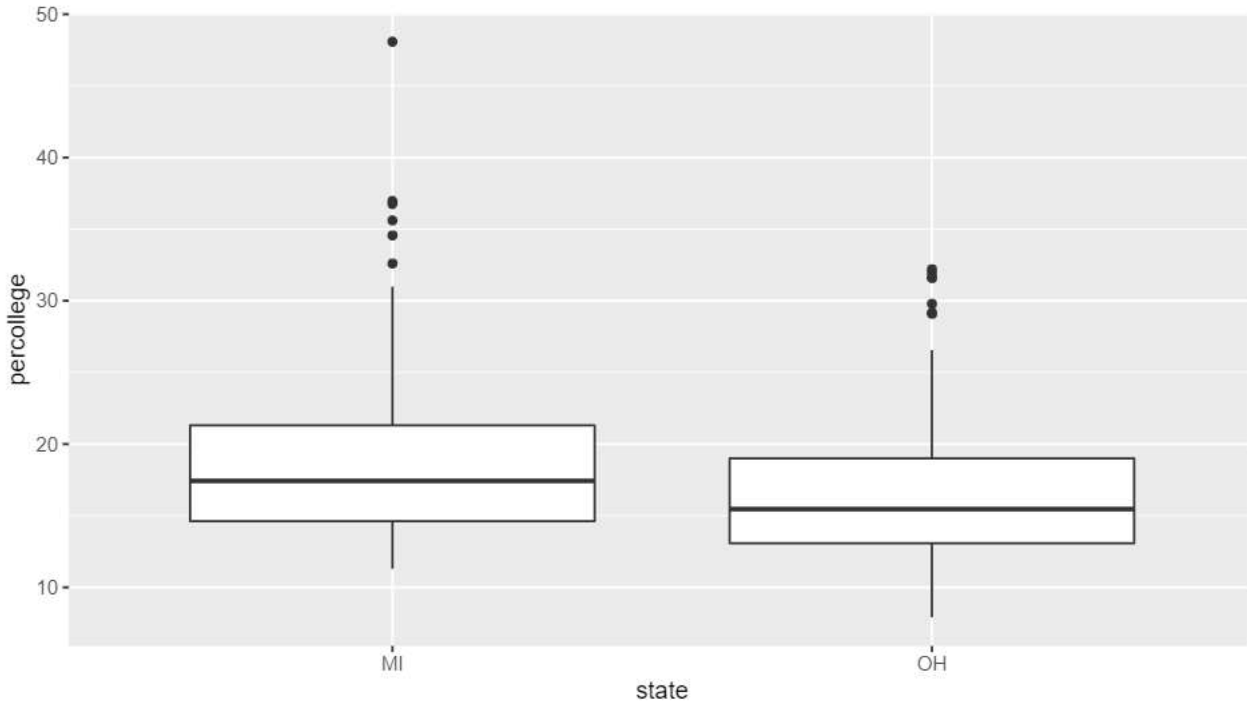
```
OH_MI <- midwest %>%
        filter(state == "OH" | state == "MI") %>%
        select(state, percollege)

# Summary statistics
summary <- OH_MI %>%
  group_by(state) %>%
  summarize(
    min = min(percollege),
    q1 = quantile(percollege, 0.25),
    median = quantile(percollege, 0.5),
    q3 = quantile(percollege, 0.75),
    max = max(percollege),
    mean = mean(percollege),
    sd = sd(percollege)
  )

summary
```

```
## # A tibble: 2 x 8
##   state   min    q1 median    q3   max  mean    sd
##   <chr> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
```

```
## 1 MI     11.3    14.6    17.4  21.3  48.1  19.4  6.85
## 2 OH      7.91   13.1    15.5  19.0  32.2  16.9  5.82
```

To further explore the data, we create side-by-side box plots.

```
ggplot(OH_MI, aes(state, percollege)) +
  geom_boxplot()
```



We can see Ohio appears to have slightly fewer college-educated adults than Michigan but the graphic doesn't tell us if it is statistically significant or not.

In this case, we will be searching for any differences between the means rather than if one is specifically less than or greater than the other. Hence we will test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$. We use a significance level of $\alpha = 0.01$ for the test.

```
t.test(percollege ~ state, data = OH_MI, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  percollege by state
## t = 2.5953, df = 161.27, p-value = 0.01032
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -0.01106105  5.07307602
## sample estimates:
## mean in group MI mean in group OH
##         19.42146         16.89045
```

The first argument in the `t.test` above may have surprised you. Note that our data was all in one data frame, but it can be grouped by the variable state to separate the data into the appropriate two populations for the hypothesis test. So rather than entering the two vectors of data as separate arguments in the `t.test`, we use the syntax `percollege ~ state` to tell R that our data comes from the `percollege` variable, but grouped by `state` to determine the two separate vectors.

The output for this hypothesis test begins with "Welch Two Sample t-test". When we ran the t-test, we did not change `var.equal` from the default `FALSE` value, so Welch's approximation for degrees of freedom was used rather than pooling the sample standard deviations and using $n + m - 2$ degrees of freedom. Because often very little is known about the population variances, many statisticians will automatically use Welch's modified t approximation (which is suggested by the fact that this is the default in R).

EXERCISE 5

State your conclusion for the hypothesis test with a significance level of $\alpha = 0.01$. Interpret your conclusion in terms of the problem.

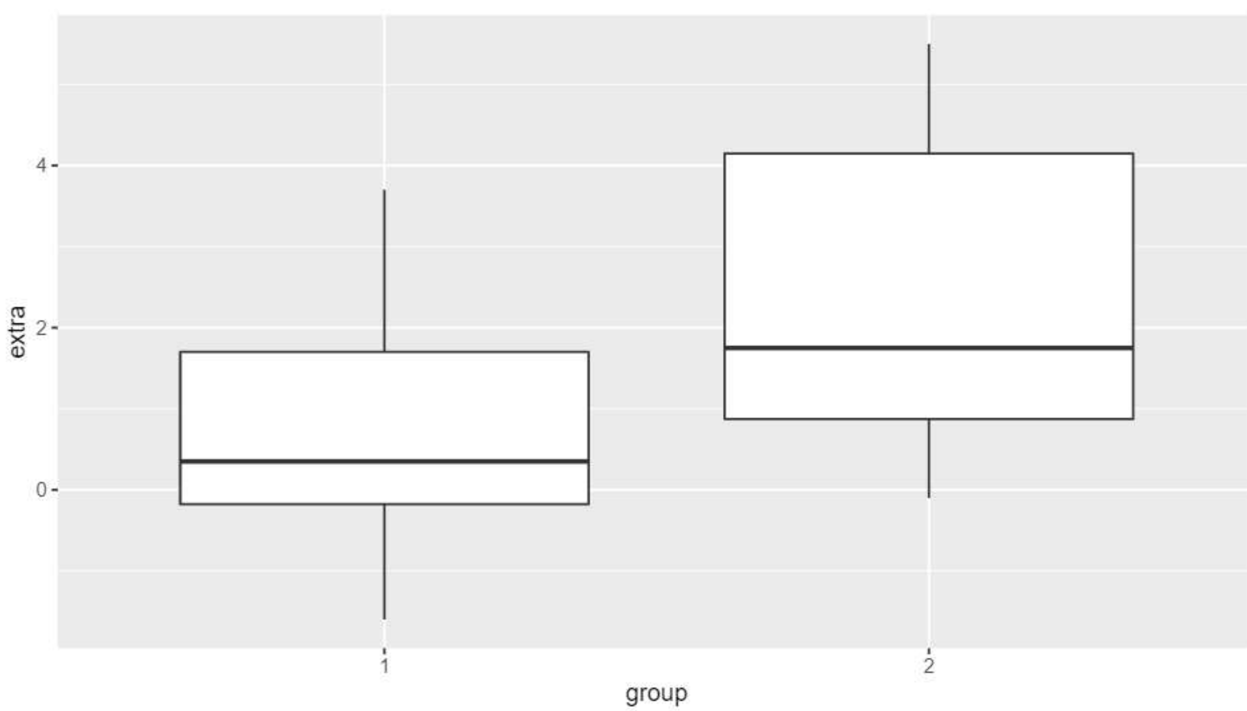
Answer: Since the p-value $= 0.01032 > 0.01 = \alpha$, we fail to reject H_0. That is with 99% confidence we cannot say that $H_1 : \mu_1 \neq \mu_2$ is true and hence we believe that the mean number of adults that are educated in Michigan is not significantly different than the mean number of adults educated in Ohio.

Paired t-test

To illustrate the paired t-test, we will analyze the built-in `sleep` data frame. Use the console to view the data frame and read the help file before continuing.

In this case, we are assessing if there is a statistically significant difference in the effect of a one particular drug on sleep compared to another drug (increase in hours of sleep compared to control) for 10 patients. We want to see if the mean values for the `extra` variable differs between group 1 (drug 1) and group 2 (drug 2). We begin by exploring the data with a side-by-side box plot.

```
ggplot(sleep, aes(group, extra)) +
  geom_boxplot()
```



We then perform the `t.test` as in the previous sections, but we add the `paired = TRUE` argument. We do not need the `alternative` argument since the default is `two.sided`, which we want here. We test the null hypothesis against a two-sided alternative hypothesis with a significance level of $\alpha = 0.01$.

```r
t.test(extra ~ group, data = sleep, paired = TRUE, conf.level = 0.99)
```

```
##
##  Paired t-test
##
## data:  extra by group
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -2.8440519 -0.3159481
## sample estimates:
## mean of the differences
##                   -1.58
```

EXERCISE 6

At a significance level of $\alpha = 0.01$, what is your conclusion for the hypothesis test of $H_0 : \mu_{Drug_1} = \mu_{Drug_2}$ versus the alternative $H_1 : \mu_{Drug_1} \neq \mu_{Drug_2}$? Interpret your conclusion in terms of the problem.

Answer: Since the p-value $= 0.002833 < 0.01 = \alpha$, we would reject H_0. That is we believe with 99% confidence that $\mu_{Drug_1} \neq \mu_{Drug_2}$.

HW Set

EXERCISE 7

According to a population census in 1986, the percentage of males who are 18 or 19 years old and are married was 3.7%. We shall test whether this percentage increased from 1986 to 1988. In a random sample of 300 males, each 18 or 19 years old, 20 were married (*U.S. Bureau of the Census, Statistical Abstract of the United States: 1988*). Write code to test this hypothesis. What is your conclusion at a significance level of $\alpha = 0.01$? Interpret your results in terms of the problem.

Answer:

```r
prop.test(20, 300, p = 0.037, conf.level = 0.99, alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  20 out of 300, null probability 0.037
## X-squared = 7.4102, df = 1, p-value = 0.003243
## alternative hypothesis: true p is greater than 0.037
## 99 percent confidence interval:
##  0.04026397 1.00000000
## sample estimates:
##          p
## 0.06666667
```

Since the p-value $= 0.003243 < 0.01 = \alpha$, we reject H_0. That is we believe with 99% confidence that the proportion of males that are 18 and 19 years old and married has not stayed the same over the last two years, more specifically we think that the proportion of married 18 and 19 year old males has increased from 1986 to 1988.

EXERCISE 8

A candy manufacturer selects mints at random from the production line and weigh them. For one week, the day shift weighed $n_1 = 194$ mints and the night shift weighed $n_2 = 162$ mints. The numbers of these mints that weighed at most 21 grams was $y_1 = 28$ for the day shift and $y_2 = 11$ for the night shift. Let p_1 and p_2 denote the proportions of mints that weight at most 21 grams for the day and night shifts, respectively. Write code to test the null hypothesis that these proportions are equal against the alternative that the proportion for the day shift is greater than the proportion for the night shift. What is your conclusion at a significance level of $\alpha = 0.05$? Interpret your results in terms of the problem.

Answer:

```
prop.test(x = c(28, 11), n = c(194, 162),alternative="greater", correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(28, 11) out of c(194, 162)
## X-squared = 5.2863, df = 1, p-value = 0.01075
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.02370925 1.00000000
## sample estimates:
##     prop 1     prop 2
## 0.14432990 0.06790123
```

Since p-value $= 0.01075 < 0.05 = \alpha$, we once again reject H_0. This means that we estimate with 95% confidence that the day proportion of selecting mints that weigh at most 21 grams does not equal the night proportion of selecting mints that weigh at most 21 grams. For non-statisticians, we believe that the proportion of mints that weigh at most 21 grams is greater in the day shift compared to the night shift.

EXERCISE 9

Let p_1 equal the proportion of 30-year-old men in Vietnam who weight over 200 pounds, and let p_2 be the proportion of 30-year-old men in the United States who weight over 200 pounds. Suppose that in a random sample of 725 Vietnamese men who were asked if they weighed more than 200 pounds, 357 responded yes, and in a random sample of 766 American men, 586 responded yes. Write code to test the hypothesis that the two population proportions are equal versus a two sided alternative. What is your conclusion at a significance level of $\alpha = 0.01$? Interpret your results in terms of the problem.

Answer:

```
prop.test(x = c(357, 586), n = c(725, 766),conf.level=0.99, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(357, 586) out of c(725, 766)
## X-squared = 119.07, df = 1, p-value < 2.2e-16
```

```
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.3346031 -0.2105954
## sample estimates:
##    prop 1    prop 2
## 0.4924138 0.7650131
```

Since the p-value = 2.2e-16 $< 0.01 = \alpha$, we once again reject H_0. This means that with 99% confidence, the proportion of Vietnamese men weighing over 200 lbs does not equal the proportion of American men weighing 200lbs or more. For our non-statisticians, we think that the proportion of Vietnamese men weighing over 200 lbs is less than the proportion of American men weighing over 200 lbs.

EXERCISE 10

Assume that the birth weight in grams of a baby born in the United States is $N(3315, \sigma^2)$. Let X equal the weight of a baby boy who is born at home in Ottawa County, and assume that the distribution of X is $N(\mu_X, \sigma_X^2)$. A random sample of 11 baby boys born at home in Ottawa County results in the following weights in grams: 4082, 3686, 4111, 3686, 3175, 4139, 3686, 3430, 3289, 3657, and 4082. Write code to input this data into R (either as a column vector or by uploading an Excel file to create a data frame) and test the null hypothesis $H_0 : \mu_X = 3315$ against the alternative hypothesis $H_1 : \mu_X > 3315$. What is your conclusion at a significance level of $\alpha = 0.05$? Interpret your results in terms of the problem.

Answer:

```
RLAB7_baby <- read_excel ("RLAB7 baby.xlsx")

t.test(RLAB7_baby, mu = 3315, alternative = "greater", conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  RLAB7_baby
## t = 4.0283, df = 10, p-value = 0.001203
## alternative hypothesis: true mean is greater than 3315
## 95 percent confidence interval:
##  3542.928      Inf
## sample estimates:
## mean of x
##  3729.364
```

Since p-value $= 0.001203 < 0.05 = \alpha$, we reject H_0. That is we expect that $\mu_x \neq 3315$ or in English: the average weight of the baby boys in grams is not equal to 3315 grams. More specifically, we think that the average weight of the baby boys is greater than 3315 grams.

EXERCISE 11

A company that manufactures motors receives reels of 10,000 terminals per reel. Before using a reel of terminals, 20 terminals are randomly selected to be tested. The test is the amount of pressure needed to pull the terminal apart from its mate. This amount of pressure should continue to increase from test to test as the terminal is "roughed up." (Since this kind of testing is destructive testing, a terminal that is tested cannot be used in a motor.) Let W equal the difference of the pressures: "test No. 1 pressure" minus "test No. 2 pressure." Assume that the distribution of W is $N(\mu_W, \sigma_W^2)$. We shall test the null hypothesis $H_0 : \mu_W = 0$ against the alternative hypothesis $H_1 : \mu_W < 0$, using 20 pairs of observations.

An excel spreadsheet with the test data is uploaded in this project, and the code below reads the data into R as a data frame. Write code to run this hypothesis test. What is your conclusion at a significance level of $\alpha = 0.05$? Interpret your results in terms of the problem.

Be sure to use the console to `View(terminals)` and understand the data frame before you begin coding!

Answer:

```
terminals <- read_excel("terminals.xlsx")
terminal_differences <- terminals %>%
  mutate(W = Test1-Test2) %>%
  select(W)
terminal_differences
```

```
## # A tibble: 20 x 1
##         W
##     <dbl>
##  1 -1.30
##  2  0.1
##  3  0.5
##  4 -1.10
##  5 -0.5
##  6  0.300
##  7  0.2
##  8  0.400
##  9 -0.600
## 10  0.100
## 11 -0.900
## 12  0.6
## 13 -0.900
## 14  0.9
## 15 -0.4
## 16 -0.300
## 17 -1.2
## 18 -0.800
## 19 -0.7
## 20 -0.2
```

```
t.test(terminal_differences, mu = 0, alternative = "less", conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  terminal_differences
## t = -1.994, df = 19, p-value = 0.03036
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##         -Inf -0.03851589
## sample estimates:
## mean of x
##     -0.29
```

Since the p-value $= 0.03036 < .05 = \alpha$, we reject H_0. That is we believe with 95% confidence that $\mu_w \neq 0$. In terms of the problem, this means that the pressure needed to pull the terminal apart differs from test 1 to test 2. More specifically, we think that the pressure needed to pull apart the terminal is greater in test 2.

EXERCISE 12A

An office furniture manufacturer installed a new adhesive application process. To compare the new process with the old process, random samples were selected from the two processes and "pull tests" were performed to determine the number of pounds of pressure that were required to pull apart the glued parts. (This kind of test is an example of destructive testing.) Let X and Y denote the pounds of pressure needed for the new and old processes, respectively. The sample observations for X and Y are included in the uploaded Excel file office.xlsx. Assuming that the populations X and Y have equal variances, write code to test the null hypothesis $H_0 : \mu_X = \mu_Y$ against the alternative hypothesis $H_1 : \mu_X > \mu_Y$ with a significance level of $\alpha = 0.05$.

Answer:

```
office <- read_excel("office.xlsx")

t.test(office$new, office$old, alternative = "greater", conf.level = 0.95)
```

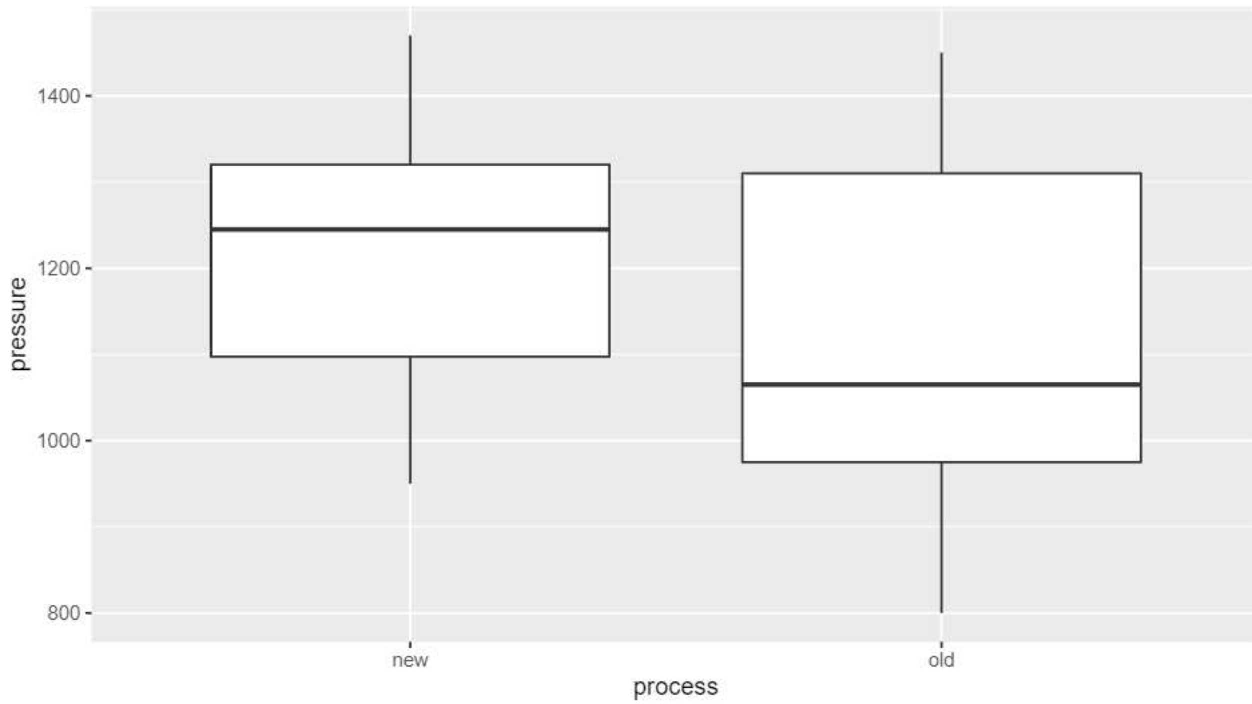
```
##
##  Welch Two Sample t-test
##
## data:  office$new and office$old
## t = 1.9566, df = 44.551, p-value = 0.02834
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  13.85157      Inf
## sample estimates:
## mean of x mean of y
##  1216.250  1118.333
```

EXERCISE 12B

What is your conclusion? Interpret your results in terms of the problem. Construct two box plots on the same graph for X and Y. Does your figure confirm your conclusion?

Answer:

```
office_new <- read_excel("office_new.xlsx")

ggplot(office_new, aes(process, pressure))+
  geom_boxplot()
```



Since p-value $= 0.02834 < 0.05 = \alpha$, we reject H_0. That is we believe with 95% confidence that $\mu_X \neq \mu_Y$. In this case we believe that the average pounds of pressure needed to pull apart the glued parts with the new process does not equal the old process. In this case we believe the new process needed more pressure to pull apart the glued parts than the old process. This is supported by our boxplot as the first quartile and the median for the old process is far below those of the new process. However, the third quartiles are relatively equal showing the variance of the old process.

EXERCISE 13A

Some measurements in mm were made on a species of spiders names *Sosippus floridanus* that are native to Florida. There are 10 females and 10 males. The lengths of their bodies are recorded in the uploaded Excel file spider.xlsx. Write code to test the null hypothesis that the lengths of female and male spiders are equal against the alternative hypothesis that female spiders are longer. Use a significance level of $\alpha = 0.025$. Do not assume that the population variances are equal. What is your conclusion? Interpret your results in terms of the problem.

Answer:

```
spider_new <- read_excel("file_show.xlsx")
t.test(spider_new$F, spider_new$M, alternative = "greater", conf.level = 0.975)
```

```
##
##  Welch Two Sample t-test
##
## data:  spider_new$F and spider_new$M
## t = 3.177, df = 11.44, p-value = 0.004207
## alternative hypothesis: true difference in means is greater than 0
## 97.5 percent confidence interval:
##  0.9611689       Inf
## sample estimates:
## mean of x mean of y
```

```
## 	15.203 	12.107
```

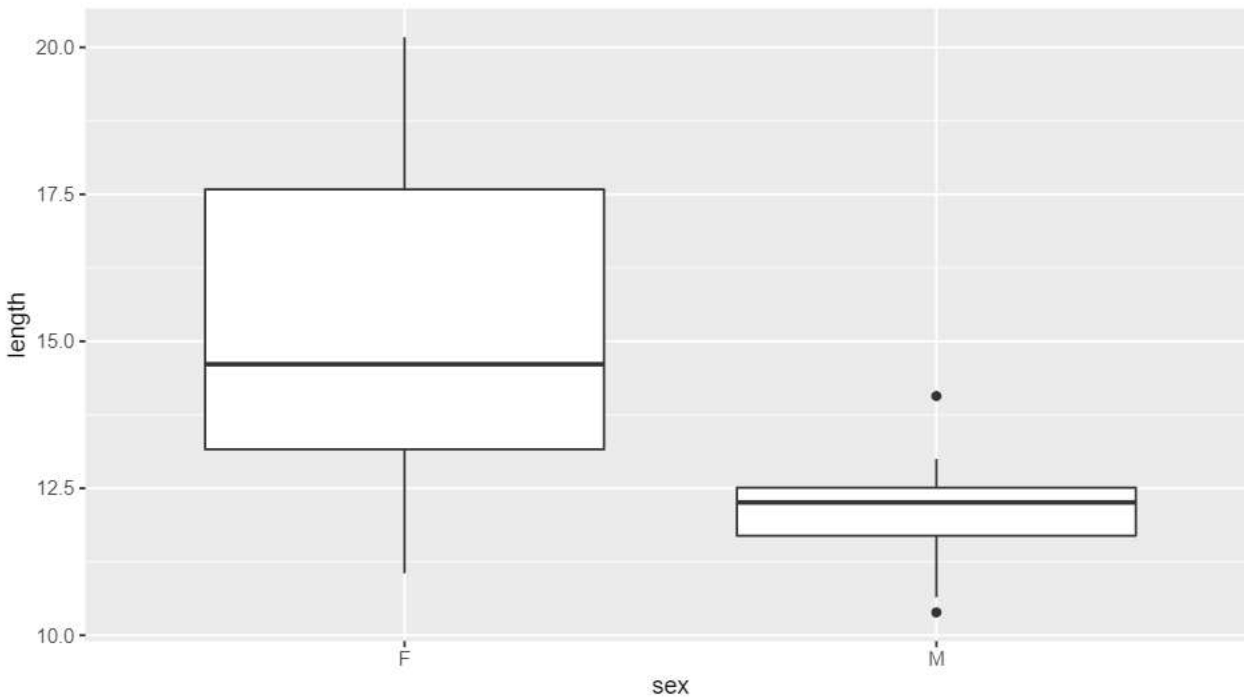
Since p-value $= 0.004207 < 0.025 = \alpha$, we reject H_0. That is we believe with 97.5% confidence that the mean length of female Sosippus floridanus does not equal the mean length of male Sosippus floridanus. For our non-statistician employer, we believe that the female Sosippus floridanus is a greater length than the male Sosippus floridanus.

EXERCISE 13B

Construct two box plots on the same graph for lengths of female and male spiders. Does your figure confirm your conclusion?

Answer:

```
spider <- read_excel("spider.xlsx")
ggplot(spider, aes(sex,length)) +
  geom_boxplot()
```



Yes our figure supports our conclusion! The male spiders appear to be much smaller than the female spider so we would reject H_0 and accept H_1.

EXERCISE 14A

Let p equal the proportion of yellow candies in a package of mixed colors. It is claimed that $p = 0.20$. To test this hypothesis, each of 20 students counted the number of yellow candies, n in a 48.1-gram package. The data is recorded in the uploaded candy.xlsx file. If each individual student tests $H_0 : p = 0.20$ against a two-sided alternative at a significance level of $\alpha = 0.05$, what proportion of the students rejected the null hypothesis?

Write code to add variables to the `candy` data frame to determine how many students rejected the null hypothesis. What proportion is this?

Answer:

```
candy <- read_excel("candy.xlsx")
```

```
candy <- read_excel("candy.xlsx")
 candy_z <- candy%>%
  mutate(z= ((y/n)-0.20)/sqrt(.2*.8*1/n))

df <- candy_z %>%
  select(z)
df
```

```
## # A tibble: 20 x 1
##          z
##      <dbl>
##  1 -1.07
##  2  0.674
##  3  0.131
##  4  0.601
##  5  0.861
##  6 -1.97
##  7  0.935
##  8  1.19
##  9  0.0680
## 10  0.674
## 11 -0.464
## 12 -1.24
## 13 -0.272
## 14  0
## 15  0.267
## 16 -0.132
## 17 -1.63
## 18 -1.51
## 19  0.131
## 20  0.788
```

```
# create new column with values 0 or 1 based on whether abs(z) > 1.96
df$abs_z <- ifelse(abs(df$z) > 1.96, 1, 0)
df
```

```
## # A tibble: 20 x 2
##          z abs_z
##      <dbl> <dbl>
##  1 -1.07       0
##  2  0.674      0
##  3  0.131      0
##  4  0.601      0
##  5  0.861      0
##  6 -1.97       1
##  7  0.935      0
##  8  1.19       0
##  9  0.0680     0
## 10  0.674      0
## 11 -0.464      0
## 12 -1.24       0
## 13 -0.272      0
```

```
## 14  0          0
## 15  0.267      0
## 16 -0.132      0
## 17 -1.63       0
## 18 -1.51       0
## 19  0.131      0
## 20  0.788      0
```

```
sum(df$abs_z)/20
```

```
## [1] 0.05
```

According to the data the proportion of students who rejected the null hypothesis is 0.05 or 1 student out of 20 students.

EXERCISE 14B

If we may assume that the null hypothesis is true, what proportion of the students would you have expected to reject the null hypothesis? Why?

Answer: 1 out of 20 because $1/20 = 0.05 = \alpha$.

EXERCISE 14C

If the 20 results are pooled so that $\sum_{i=1}^{20} y_i$ equals the number of yellow candies and $\sum_{i=1}^{20} n_i$ equals the total sample size, do you reject $H_0 : p = 0.20$? Write code to determine these sums and test the hypothesis. What is your conclusion?

Answer:

```
yy<-sum(candy$y)
nn<-sum(candy$n)

prop.test(yy, nn, p =0.20, conf.level=0.95, alternative ="two.sided", correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  yy out of nn, null probability 0.2
## X-squared = 0.18706, df = 1, p-value = 0.6654
## alternative hypothesis: true p is not equal to 0.2
## 95 percent confidence interval:
##  0.1727403 0.2190181
## sample estimates:
##         p
## 0.1948399
```

Since p-value $= 0.6654 > 0.05 = \alpha$, we fail to reject H_0. That is, we cannot say with 95% confidence that $H_1 : p \neq 0.20$ is true. Hence we suspect that $p = 0.20$ is likely true.