Confidence Interval of Difference

Introduction

In this unit, we learned that, in the case of two-treatment experiments, we can test whether the populations that receive each treatment are different from each other by testing whether their difference is different from zero. While we will often use the **t-test** to calculate the probability the two populations are equal, given their measured values, there are times when we may want to report the confidence interval around their difference.

These confidence intervals also be can be gained from the `t.test()` function, but calculating them manually in these exercises is a good opportunity to sharpen your R skills and hopefully will give you are greater conceptual sense of how the confidence interval works.

To create the confidence interval for the difference between two populations, there are the following steps:

1. Calculate the differences between the populations in each pair or block.
2. Calculate the overall mean difference between the two populations.
3. Calculate the standard deviation and standard error of the difference.
4. Look up the appropriate t-value based on the desired level of confidence and degrees of freedom
5. Add and subtract the product of standard error and t-value from the mean difference.

Example 1

We will work with the soybean_fungicide data from the previous exercise:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
soybean = read.csv("data/soybean_fungicide.csv")
head(soybean)
```

```
##   Block Treatment Yield
## 1     1         A  71.5
## 2     1         B  75.1
## 3     2         A  73.5
## 4     2         B  77.7
## 5     3         B  77.7
## 6     3         A  72.4
```

Calculate Differences

We will use the `spread()` and `mutate()` functions from the previous exercises to calculate a the population difference within each block.

```
soybean_wide = soybean %>%
  spread(Treatment, Yield)

soybean_diff = soybean_wide %>%
  mutate(diff = A-B)

soybean_diff
```

```
##    Block    A    B diff
## 1      1 71.5 75.1 -3.6
## 2      2 73.5 77.7 -4.2
## 3      3 72.4 77.7 -5.3
## 4      4 75.0 78.2 -3.2
## 5      5 76.4 75.9  0.5
## 6      6 76.5 77.5 -1.0
## 7      7 74.8 80.7 -5.9
## 8      8 75.6 78.4 -2.8
## 9      9 73.2 80.7 -7.5
## 10    10 74.6 76.9 -2.3
```

Calculate the Mean Difference

This is the mean for the new column, *diff*. As a reminder, `soybean_diff$diff` tells R to use the diff column of the soybean_diff data frame.

```
diff_mean = mean(soybean_diff$diff)
diff_mean
```

```
## [1] -3.53
```

Calculate the Standard Deviation and Standard Error

We can calculte the standard deviation using the `sd()` function.

```
diff_sd = sd(soybean_diff$diff)
diff_sd
```

```
## [1] 2.351383
```

The standard error is the standard deviation, divided by the square root of the number of observations, n. In this example, n=10. We can verify this using the `length()` function of R. We can then calculate the standard error, which we will call *diff_se*.

```
diff_N = length(soybean_diff$diff)

diff_se = diff_sd / sqrt(diff_N)
```

diff_se is the standard error. The formula we used is equal to 2.35 / sqrt(10). We just used the R objects to which those values were assigned, to save retyping.

Calculate the t-value

We calculate the t-value to use in our confidence interval by using the `qt()` function of R. Remember, the qt function uses two arguments. The first is the desired level of confidence. We want a 95% confidence interval, so we will use 0.975. That will leave 0.025, or 2.5%, beyond the upper and lower limits of our confidence interval.

The second argument is the degrees of freedom. Since we have observed ten differences, we have nine degrees of freedom.

```
t_value = qt(0.975, 9)
t_value
```

```
## [1] 2.262157
```

Our t_value is approximately 2.26.

Calculate the Confidence Interval

The confidence interval is bound by **upper** and **lower confidence limits**. The lower limit is equal to the mean difference minus the product of the standard error of the difference and t-value. The upper limit is the mean difference plus the product of the standard error of the difference.

```
lower_limit = diff_mean - (diff_se * t_value)
upper_limit = diff_mean + (diff_se * t_value)

lower_limit
```

```
## [1] -5.212078
```

```
upper_limit
```

```
## [1] -1.847922
```

The 95% confidence interval, with 9 df, for the difference is (-5.21, -1.85). Note the confidence interval does not include zero, therefore the two populations (and, thus, the treatments they received) are significantly different at the $P<0.05$ level.

Which treatment was greater? Since we subtracted B from A, the negative confidence interval indicates B was greater.

Example 2

We will work with the Darwin corn data.

```
darwin = read.csv("data/darwin.csv")
head(darwin)
```

```
##   pot pair  type height
## 1   I    a cross 23.500
## 2   I    a  self 17.375
## 3   I    b cross 12.000
## 4   I    b  self 20.375
## 5   I    c cross 21.000
## 6   I    c  self 20.000
```

Calcuate Differences

```
darwin_wide = darwin %>%
  spread(type, height)
head(darwin_wide)
```

```
##   pot pair  cross   self
## 1   I    a 23.500 17.375
## 2   I    b 12.000 20.375
## 3   I    c 21.000 20.000
## 4  II    d 22.000 20.000
## 5  II    e 19.125 18.375
## 6  II    f 21.500 18.625
```

```
darwin_diff = darwin_wide %>%
  mutate(diff = cross - self)
head(darwin_diff)
```

```
##   pot pair  cross   self   diff
## 1   I    a 23.500 17.375  6.125
## 2   I    b 12.000 20.375 -8.375
## 3   I    c 21.000 20.000  1.000
## 4  II    d 22.000 20.000  2.000
## 5  II    e 19.125 18.375  0.750
## 6  II    f 21.500 18.625  2.875
```

Calculate Mean Difference

```
darwin_diff_mean = mean(darwin_diff$diff)
darwin_diff_mean
```

```
## [1] 2.616667
```

Calculate Standard Deviation and Standard Error

```
darwin_diff_sd = sd(darwin_diff$diff)
darwin_diff_N = length(darwin_diff$diff)

darwin_diff_se = darwin_diff_sd / sqrt(darwin_diff_N)
darwin_diff_se
```

```
## [1] 1.218195
```

Calculate the t-value

There were 15 pairs, so there are 14 degrees of freedom. We will again calculate the t-value for the 95% confidence interval.

```
darwin_t_value = qt(0.975, 14)
darwin_t_value
```

```
## [1] 2.144787
```

Calculate the Confidence Interval

```
darwin_lower_limit = darwin_diff_mean - (darwin_diff_se * darwin_t_value)
darwin_upper_limit = darwin_diff_mean + (darwin_diff_se * darwin_t_value)

darwin_lower_limit
```

```
## [1] 0.003899165
```

```
darwin_upper_limit
```

```
## [1] 5.229434
```

The confidence interval was (0.004, 5.229). It does not include zero, so the difference is significant. Since we subtracted the self-pollinated values from the cross-pollinated values, the positive difference means the cross-pollinated plants were taller than the self-pollinated plants.