Protist Lab: t-test

Sam Bogan

Summary

The *t*-test is a simple but powerful statistical tool for evaluating differences between values measured in two groups. Like all statistical tests, it relies on assumptions about the data you are using and becomes more reliable with larger sample sizes. This walk through will describe the basic structure and application of the *t*-test before helping you run R code to (i) plot your data for visual inspection of differences in group means, (ii) testing whether your data match the *t*-test's assumptions, and (iii) applying and analyzing the results of a *t*-test.

What you will need

i. The 't-test' folder that this .rmd script is stored in and all of its original contents

ii. R and R studio

What is a *t*-test?

Say you want to pick apples from two farms, one farm with lots of water and sunshine and another in a cloudy place with poor irrigation. Since you're an Intro Bio student, you're *obviously* going to these farms with a clear and testable hypothesis in mind; if water and sunshine benefit apple growth, the farm with more sunshine and water should produce larger apples. So, you go to the farms, collect five apples from each one, and measure the mass of the apples. A *t*-test will tell you the probability that the two groups of apples you measured could have come from a single population (i.e., the same farm with the same sunshine and water). This probability is called the ***p*-value: the probability that the values of two groups could be randomly drawn from a single distribution**. In simple terms, the p-value is the probability that there is no true difference in means between groups. | After weighing your apples, you see that there is a noticeable difference in mean mass between farms, but your *t*-test demonstrates that there is a high probability these apples could have been drawn from the same population ($p = 0.6$; 60% chance that the measurements were drawn from a single distribution). Being the meticulous scientist you are, you recognize that your sample size may not be large enough to detect differences between farms. You go back to both farms and measure the mass of 95 more apples from each, bringing your sample size to 100 per farm. You reapply your *t*-test to these new and improved data and *voila*, a second *t*-test determines that the mean difference between farms is unlikely to be drawn from a single distribution ($p = 0.001$; 0.1% chance). Similar to the effect of sample size, greater variance or standard deviation in mass of a farm's apples will also increase your estimated *p*-value; if the distribution of apple mass at the two farms overlaps more, it is more likely that these two means could have been derived from a single distribution.

t-tests also generate a statistical parameter called t, the test's namesake. t **represents the ratio of differences in a value between groups over the total variation of that value, accounting for sample size**. When the absolute value of t is higher, discernable differences between group means are greater. An equation for calculating t is illustrated below. This equation corresponds to the Student's

t-test. We will be using Welch's t-test in this walkthrough, but the equation for Student's t is more easily interpreted and is included for illustrative purposes:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{(s^2(\frac{1}{n_1} + \frac{1}{n_2}))}}$$

t = t statistic

$\overline{x_i}$ = mean of groups 1 or 2

s = standard deviation of all observations

n_i = sample size of groups 1 or 2

The t-test assumes normality of data

t-tests assume several things about the nature of your data. One of these assumptions is that your data are normally distributed, meaning the histogram of frequency for K values exhibits a bell shaped curve with a mean and equal standard deviations in the postive and negative directions from that mean. Below is an example of a normally distributed histogram for a hypothetical value called 'x' (Fig. A).

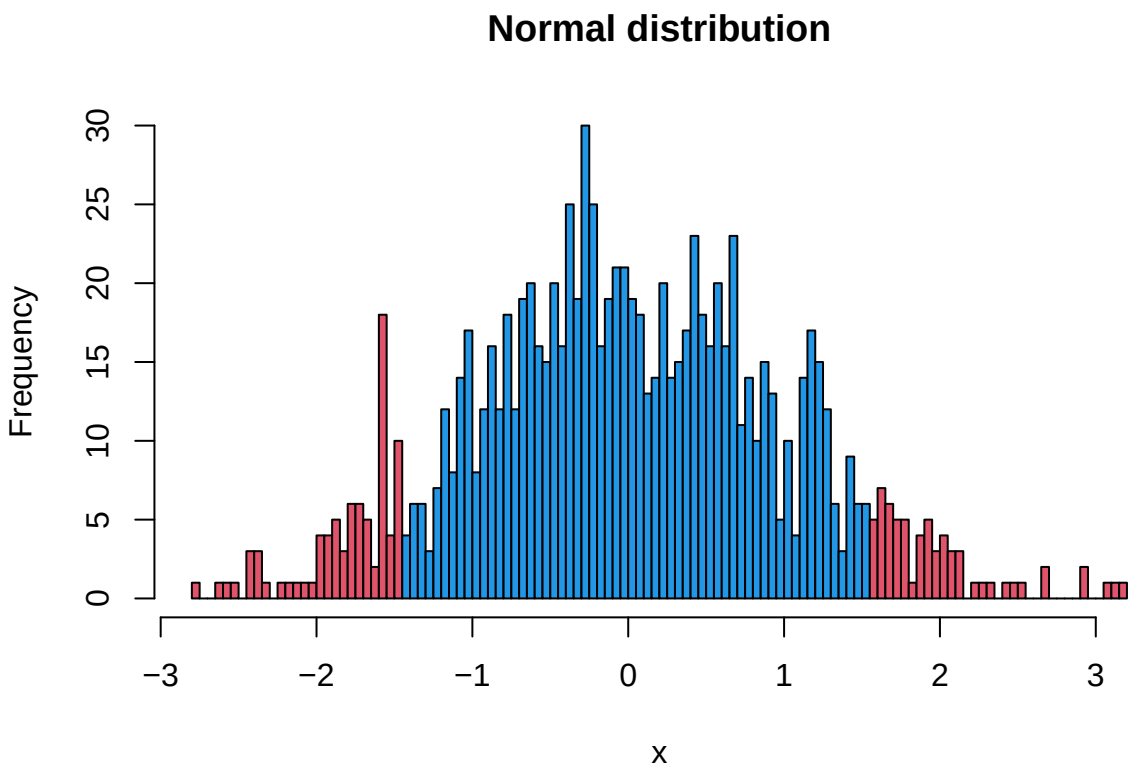

Figure A | A histogram of simulated normal data. Blue bars depidct ranges of data within the upper and lower quartiles of x. Red bars depict extreme ends or 'tails' of the normal distribution. One sign that these data are normally distributed is that the positive and negative tails are of equivalent ranges; extreme values are not skewed in one direction or the other.

Not all data that you analyze will be normally distributed. Below is a histogram of simulated data that exhibit a skewed, non-normal distribution. These data would not meet the assumption of normality necessary to conduct a t-test (Fig. B).

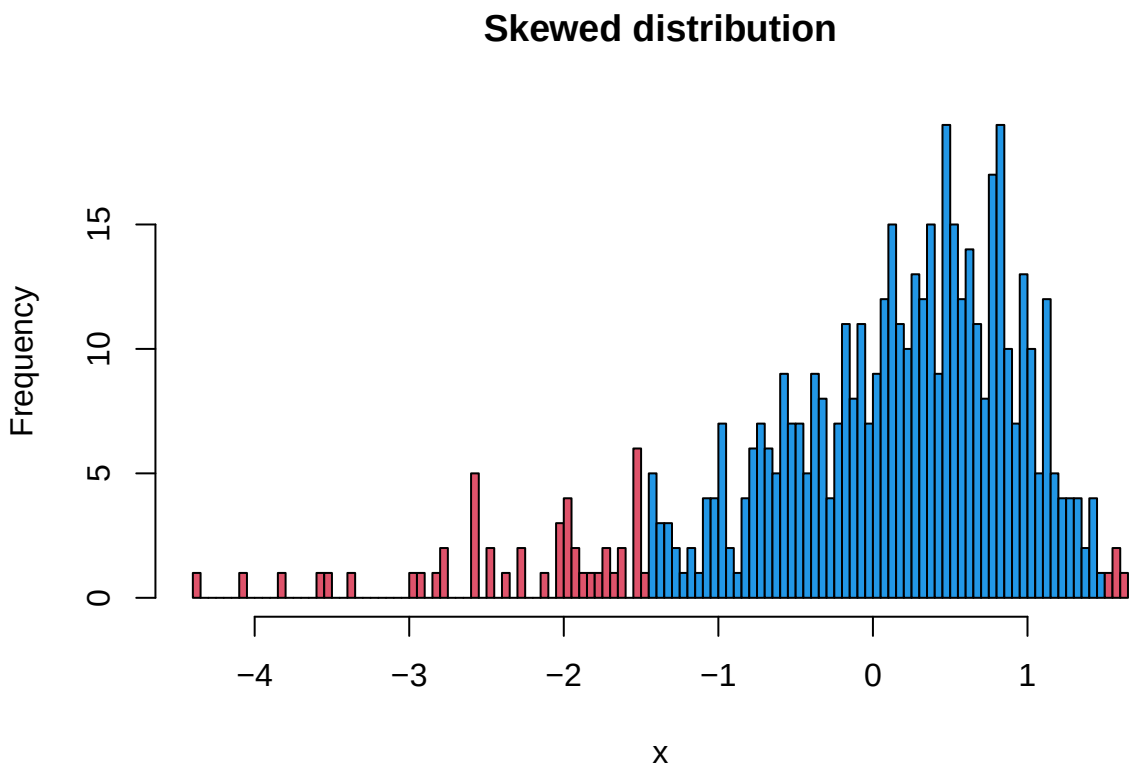

Figure B | A histogram of simulated non-normal data. Blue bars depidct ranges of data within the upper and lower quartiles of x. Red bars depict extreme ends or 'tails' of the normal distribution. As these data are postiviely skewed, the negative tail of the distribution exhibits a much wider range than the positive distribution.

Before running a t-test, make sure to look at the distribution of your data and confirm that it is normal! Empirical tests of normality also exist, but we will not cover these here.

We have generated some nifty data that lend themselves to a t-test. We need to compare K (carrying capacity) estimates between groups of protists cultured with and without competitors to see whether and how competition affects limits on population size. To apply a t-test to these data, let's first import it into R using the code chunk below. In these code chunks, the '#' symbol denotes a description of what each line of code is doing. For the purposes of this lab however, it is more important that you understand the code's output rather than the code itself.

Import dataset to R

```
# Read .csv file containing protist K data into R
Protist_K_Data <- read.csv("Protist_K_Data.csv")
```

```r
# Let's look at how this data sheet is formatted
print(Protist_K_Data)
```

```
##    A_alone_K B_alone_K A_mixed_K B_mixed_K
## 1       3500      3680      1000      3000
## 2       3400      3680       800      3050
## 3       3300      3750      1300      3100
## 4       3650      3910      1150      3200
## 5       3600      3730      1050      2850
## 6       3550      3580       950      3050
## 7       3600      3660      1100      2950
## 8       3400      3730       800      2900
## 9       3350      3500      1100      2900
## 10      3450      3680       850      3050
## 11      3550      3910       750      3000
## 12      3650      3530      1050      2850
## 13      3550      3500       700      2800
## 14      3300      3830      1000      3200
## 15      3950      3710       750      3250
## 16      3450      3250       950      3250
## 17      3400      3680      1050      3200
## 18      3500      3750      1000      3000
```

Looking at the 'Protist_K_Data' data sheet, we can see that there are 4 columns. Each column has a short but unique name. We will use these names to identify groups that we want to compare using the t-test. Next, we will plot and visually compare our K measurements between each group. This will allow us to identify groups that we may want to compare.

Plotting data

```r
## Don't worry about any of the code in this chunk
## Just look at the graph that it creates (below)

# Reformat datasheet for plotting
Plotting_Data <- rbind( data.frame(Species = "A",
                                   Mixed = "Alone",
                                   K = Protist_K_Data$A_alone_K),
                        data.frame(Species = "B",
                                   Mixed = "Alone",
                                   K = Protist_K_Data$B_alone_K),
                        data.frame(Species = "A",
                                   Mixed = "Mixed",
                                   K = Protist_K_Data$A_mixed_K),
                        data.frame(Species = "B",
                                   Mixed = "Mixed",
                                   K = Protist_K_Data$B_mixed_K))

# Create a variable combining species and treatment
Plotting_Data$Group <- as.factor(paste(Plotting_Data$Species,
                                       Plotting_Data$Mixed,
```

```
                                        sep = " "))

# Plot K values across all four treatment groups
plot(Plotting_Data$K ~ Plotting_Data$Group, main = "K across species and treatments",
     xlab = "Treatment", ylab = "K", pch = 19)
```

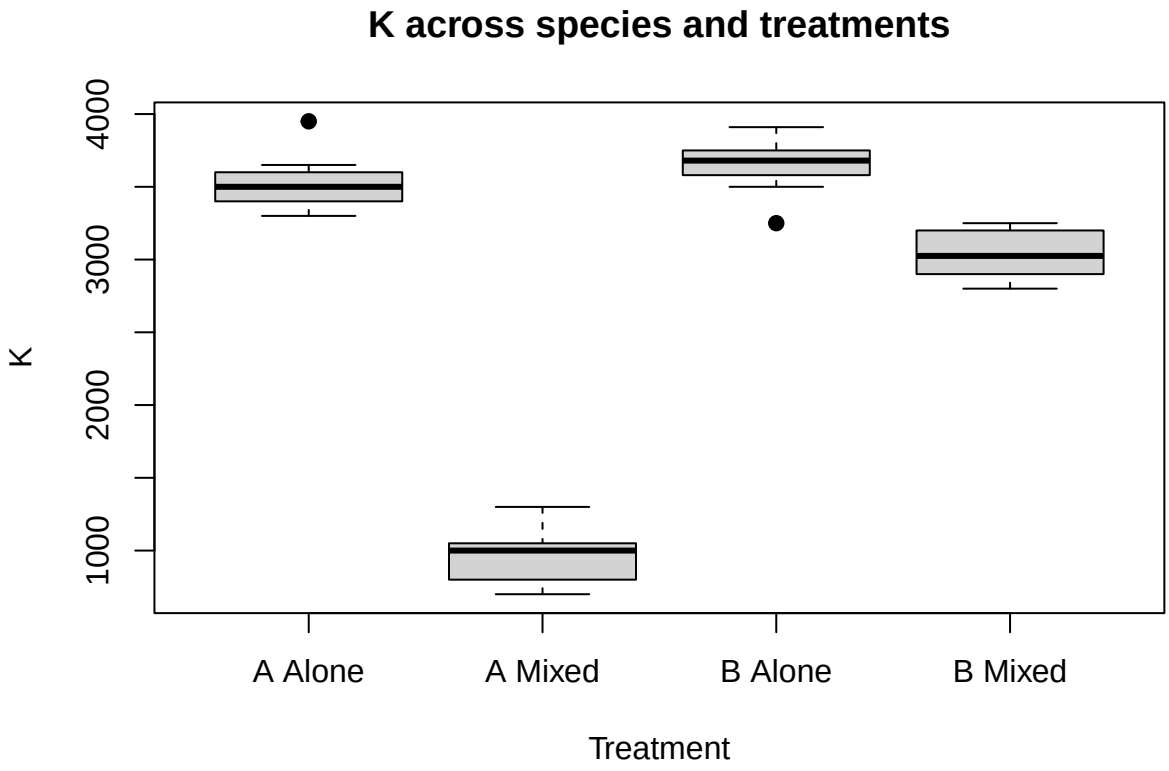

Figure 1 | A box-whisker plot of K estimates for species and treatments. Dark black lines depict median K. Boxes depict upper and lower quartiles. Error bars depict the minima and maxima of K distributions. Points depict outliers.

It looks like there are some clear differences among treatment groups. For example, competition seems to have a large effect on K in both A and B, with A exhibiting the greatest reduction in K under the mixed treatment (Fig. 1). We will perform a few t-tests to see if K is significantly different between these groups. First, we must check to see if our data meet the assumptions of a t-test.

Let's take a look at the distribution of our own data. Since there are noticeable differences between species and treatment groups, a histogram of K is not going to look normal (it will look like more like the ridgeline of a mountain). We can get around this issue by 'scaling' our data, where we set the mean of each group to 0 and transform K into standard deviations of the mean. This may sound complicated, but it is simply a way of looking at the shape of a total distribution across groups with different means. After doing so, we can see that K appears to be normally distributed across all groups (Fig. 2).

```
# Scale K data
K_scaled <- rbind( scale( Protist_K_Data$A_alone_K),
       scale( Protist_K_Data$B_alone_K),
       scale(Protist_K_Data$A_mixed_K),
       scale(Protist_K_Data$B_mixed_K))
```

```r
#Plot scaled K as a histogram
hist(K_scaled,
     main = "Distribution of K",
     xlab = "K (scaled)")
```

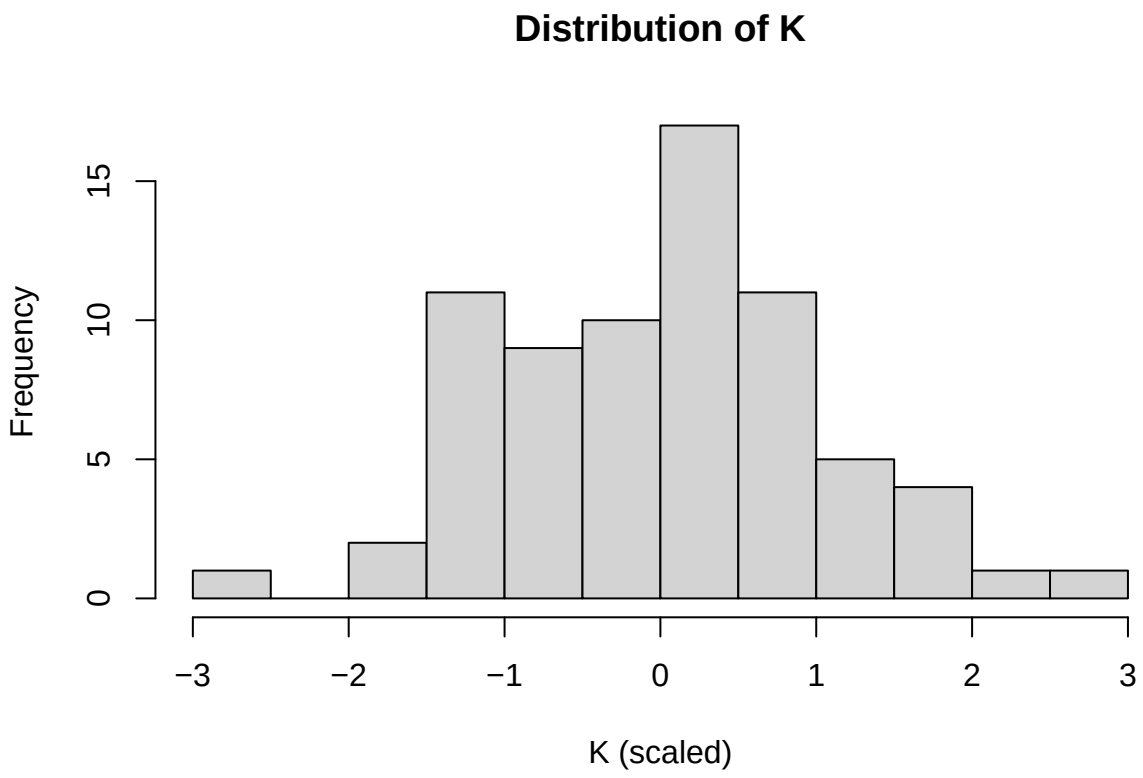

Figure 2 | A histogram of scaled K values from all species and treatment groups, meaning data were transformed such as that species and treatment group means = 0 and K is expressed in standard deviations from the mean.

Looking at our data in this manner is great, and helps catch bugs or other issues. However, we can go beyond visual inspection and apply formal tests of normality. This cannot be done using raw data or scaled data. Rather, formal tests of normality rely on what are called "residuals": variation in a measurement that remains after explaining differences between groups do to independent variables (e.g., species or treatment group). We won't conduct a normality test in this walk through but if you are interested in learning about the method, please check out this excellent resource: https://www.datanovia.com/en/lessons/normality-test-in-r/.

Performing the t-test

Now that we know our data are normally distributed, it's time to compare species and treatment groups using the t-test. We will use R's 't.test' function using a two sided test. Two sided tests allow us to ask "Is the mean of one group significantly different from the other?". By contrast, a one sided test would ask "Is the mean of this group significantly greater than the other?". The output of a two sided t-test contains several metrics, two of which are t and p.

p, as described earlier, is the p-value: the probability that observations in two groups could have been drawn from a single distribution. Follow along with the code chunks below to see how we take our data sheet (Protist_K_Data), input it to the 't.test' function, and interpret our results.

Comparing K of species cultured alone

First we will apply a two sided t-test to K values of species A vs. B when grown alone. In plain terms, this asks "Do populations of species A and B have different carrying capacities?"

```
# Two sided t-test: species A vs B cultured alone
A_vs_B_alone <- t.test(x = Protist_K_Data$A_alone_K,
                       y = Protist_K_Data$B_alone_K,
                       alternative = "two.sided",
                       mu = 0, paired = FALSE, var.equal = FALSE,
                       conf.level = 0.95)

# Print results
A_vs_B_alone
```

```
##
##  Welch Two Sample t-test
##
## data:  Protist_K_Data$A_alone_K and Protist_K_Data$B_alone_K
## t = -3.0929, df = 33.991, p-value = 0.003946
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -267.89383  -55.43951
## sample estimates:
## mean of x mean of y
##  3508.333  3670.000
```

As you can see, **differences between K_A and K_B when cultured under unmixed conditions (K_{A_U} and K_{B_U}) correspond to a t-value of -3.09 and a p-value of 0.004**, meaning there is a significant difference in K between species when cultured alone with a 0.4% chance that these differences could have been randomly sampled from a single distribution.

##Comparing K of species A in mixed and unmixed cultures | We can also apply a t-test to compare the distributions of K_{A_M} and K_{A_U}, the carrying capacities of species A under mixed and unmixed conditions.

```
# Two sided t-test: species A alone vs. mixed
Mixed_vs_Alone_A <- t.test(x = Protist_K_Data$A_alone_K,
                       y = Protist_K_Data$A_mixed_K,
                       alternative = "two.sided",
                       mu = 0, paired = FALSE, var.equal = FALSE,
                       conf.level = 0.95)

# Print results
Mixed_vs_Alone_A
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  Protist_K_Data$A_alone_K and Protist_K_Data$A_mixed_K
## t = 48.27, df = 33.964, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2437.314 2651.575
## sample estimates:
## mean of x mean of y
## 3508.3333  963.8889
```

##Comparing K of species B in mixed and unmixed cultures | **Differences in K_{A_M} and K_{A_U} are much larger than what we observed between K_{A_U} and K_{B_U} (t = 48.27)**. Furthermore, there is almost no chance that our observations of K_{A_M} and K_{A_U} were drawn from the sample population/distribution (p < 2.376e-14). If we hypothesized that the carrying capacity of species A is influenced by competition with species B, these results provide strong support for rejecting our null hypothesis.

```
# Two sided t-test: species B alone vs. mixed
Mixed_vs_Alone_B <- t.test(x = Protist_K_Data$B_alone_K,
                        y = Protist_K_Data$B_mixed_K,
                        alternative = "two.sided",
                        mu = 0, paired = FALSE, var.equal = FALSE,
                        conf.level = 0.95)

# Print results
Mixed_vs_Alone_B
```

```
##
##  Welch Two Sample t-test
##
## data:  Protist_K_Data$B_alone_K and Protist_K_Data$B_mixed_K
## t = 12.653, df = 33.686, p-value = 2.376e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  534.3705 738.9628
## sample estimates:
## mean of x mean of y
##  3670.000  3033.333
```

Differences in K_{B_M} and K_{B_U} are also larger than what we observed between K_{A_U} and K_{B_U} (t = 12.65). Furthermore, there is almost no chance that our observations of K_{B_M} and K_{B_U} were drawn from the sample population/distribution (p = 2.38e-14). If we hypothesized that the carrying capacity of species B is influenced by competition with species A, these results provide strong support for rejecting our null hypothesis.

Lastly, let's compare K of A and B in mixed treatments (K_{A_M} vs. K_{B_M}). We'll ask whether the difference (t) between the species under mixed conditions is greater than their difference in K when cultured unmixed If this is the case, t associated with K_{A_M} vs. K_{B_M} should be greater than t associated with K_{A_U} vs. K_{B_U}.

```
# Two sided t-test: species A alone vs. mixed
A_vs_B_mixed <- t.test(x = Protist_K_Data$A_mixed_K,
                        y = Protist_K_Data$B_mixed_K,
                        alternative = "two.sided",
                        mu = 0, paired = FALSE, var.equal = FALSE,
```

```
                       conf.level = 0.95)

# Print results
A_vs_B_mixed
```

```
##
##  Welch Two Sample t-test
##
## data:  Protist_K_Data$A_mixed_K and Protist_K_Data$B_mixed_K
## t = -40.754, df = 33.573, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2172.689 -1966.200
## sample estimates:
## mean of x mean of y
##  963.8889 3033.3333
```

```
# Fold-change difference: t of A and B mixed / t of A and B alone
t_foldchange <- A_vs_B_mixed$statistic / A_vs_B_alone$statistic

# Print t_foldchange
t_foldchange
```

```
##       t
## 13.1766
```

By comparing K_{A_M} vs. K_{B_M}, we see a significant t equal to -40.754. The difference in K between species when mixed was 13.18x greater than t of K in unmixed cultures; The carrying capacity of Species A was more reduced under competition than that of Species B.