Session 6

Drs. Ibraheem Ali and Michael Wells

2023-03-02

# Introduction to Statistical Models

In session 2 we described a number of fundamental concepts in statistics. We discussed how statistics can be used to make inferences about a population based on a random sample. We discussed summary statistics that can be used to describe characteristics of a population’s center and spread. We introduced hypothesis testing used to compare between two different samples. In session 4 we described visualization tools that can be used in conjunction with statistics to better understand features of samples that you collect.

In this session, we will cover more of why and how to use specific types of statistical tests. Ultimately in your experimental journey, you will collect data which you will want to characterize and compare with different conditions. Sometimes you are comparing if one group is different from another. Other times you might be assessing relationships between variables - for example how one gene expression changes in response to the quantity of expression of another gene. Other times you will be looking for consistencies among groups of many variables as you might do during gene pathway analysis.

By the end of this session you should be able to:

* Recognize when to use a t-test, how to run the test in R, and how to interpret the results
* Recognize when to use an analysis of variance (ANOVA) test versus a t-test
* Distinguish between one-way and two-way ANOVA test
* Determine which statistical model is best for explaining a data set
* Run a Dunnett’s test for multiple comparison

## Loading Packages

For our lesson, we will need to do some data wrangling and data visualization in conjunction with our statistical tests. Notice which packages are listed below and recall what the packages were used for. Many of the most common statistical tests are found in the stats package which often comes pre-loaded in the system library, but load the library just to be sure. You may also notice the package pheatmap. This package is especially useful when constructing “pretty heatmaps” that go beyond the capability than what ggplot2 can offer.

rm(list=ls())  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.1 ✔ purrr 1.0.1  
## ✔ tibble 3.1.8 ✔ dplyr 1.1.0  
## ✔ tidyr 1.3.0 ✔ stringr 1.5.0  
## ✔ readr 2.1.3 ✔ forcats 0.5.2  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(ggplot2)  
library(ggExtra)  
library(RColorBrewer)  
library(stats)  
library(pheatmap)  
library(DescTools) # for convenient statistical summaries  
library(AICcmodavg) # for comparing statistical models

## Background

Cerebral organoids are a powerful human stem cell-based three- dimensional model of early brain development and associated diseases like autism and intellectual disability. You are studying a gene called CACHD1, which was recently identified as a regulator of neural proliferation and differentiation in two-dimensional models. You generate cell lines that harbor either CRISPR-mediated mutations in CACHD1 or non-targeting controls. You then assess organoid size/growth to determine if disrupting this gene impacts organoid development.

## Importing Data

OrganoidSize\_Mutant <- read.csv("../Datasets/CACHD10OrgSize.csv", header = T, sep = ",")  
  
OrganoidSize\_KD <- read.csv("../Datasets/cachd1-tbr1-kd.csv", header = T, sep = ",")

## Statistical Models

Statistical models provide a framework of assumptions about populations in order to make comparisons between samples. Different statistical tests use different statistical models in order to calculate the probability that a given hypothesis is true (or not true).

# T-testing

A *t-test* is a statistical test that is used to compare the means of two groups. It is typically used to determine if a process or treatment has an effect in the population of interest. The test uses characteristics of the mean and standard deviation to test the probability that a null or alternative hypothesis is true. In a typical two-sided t-test, these are the following hypotheses that are tested:

* The null hypothesis is that the true difference between the two groups is 0
* The alternative hypothesis is that the true difference is not equal to 0

Mathematically, a t-test calculates the “t” value which is equal to the difference between the sample means divided by the “pooled” standard error which factors in both standard deviation and the number of samples used to calculate each mean. The **t-value** measures the size of the difference in means for two samples relative to the variation in the sample data. The larger the t-value, the lower the probability that your null hypothesis is correct. In other words, the higher the t-value, the lower the p-value.

T-tests, and a number of other statistical tests, factor in a value called **“degrees of freedom”**, which refers to the maximum number of logically independent values in a sample. The more values collected to calculate the mean, the higher the number of degrees of freedom. Degrees of freedom are used to characterize specific distributions and are necessary for many tests including t-tests, F-tests and chi-square tests. The number of degrees of freedom paired with the t-value to calculate the appropriate probability based on the expected sample distribution.

There are a number of different types of t-tests that can be used each with a slightly different set of assumptions. A basic student’s T-test assumes that the two populations being studied are *independent* from one another and have equal variance. A Welch’s T-test assumes that the two populations are normally distributed, but have unequal population variances. **It is generally recommended to use a Welch’s T-test to compare between samples** because it creates a more accurate probability if samples have unequal variances, but is equally accurate to a Student’s t-test when the samples have equal variance.

As we mentioned above T-tests can be one or two sided. Two sided t-tests test if one sample **is different** from another sample. A one-sided t-test is used to check the direction in a difference in means. In other words, a one-sided test can be used to see if one sample’s mean is **greater than**, or **less than** another sample mean. If you are using a one sided test, your null hypothesis stays the same, but the alternative hypothesis changes depending on the direction you are testing. So instead of: “the alternative hypothesis is that the true difference is **not equal to** 0”; it is “the alternative hypothesis is that the true difference is **greater than**/**less than** 0”.

## Welch’s t-test

During experimentation, the authors initially found that after 27 days of development, organoids from the control sample appeared visually smaller than those from the mutant sample. Using the OrganoidSize data: filter out all values that were not collected on day 27. Create one data frame for the control, and another for CACHD1\_Mutant1. Use the values in the “Size” column to run a one-sided t-test to see if the samples are significantly different.

OrganoidSizeDay27 <- OrganoidSize\_Mutant %>%  
 filter(Day == 27)  
  
ControlSize <- OrganoidSizeDay27 %>%  
 filter(Condition == "CONTROL")  
  
Mutant1Size <- OrganoidSizeDay27 %>%  
 filter(Condition == "CACHD1\_MUTANT1")

The t.test() command will test if the CACGH1 knockout is significantly larger compared to the control sample. In this case since we are comparing the first sample with the second, our alternative hypothesis would be that the true difference in means is **less** than 0, or a negative value. In this command, the first and second arguments must be a vector of numeric (or logical) values. The third argument we need to give is what to set the alternative hypothesis. There are three options

* “greater” - tests the hypothesis that x is greater than y
* “less” - tests the hypothesis that x is less than y
* “two.sided” - tests the hypothesis that x is equal to y (default)

Since the default is “two.sided” we must designate alternative = "less" in order for R to know to do a one-sided test. See the example below.

t.test(ControlSize$Size, Mutant1Size$Size, alternative = "less")

##   
## Welch Two Sample t-test  
##   
## data: ControlSize$Size and Mutant1Size$Size  
## t = -2.4484, df = 27.756, p-value = 0.01047  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -0.04779575  
## sample estimates:  
## mean of x mean of y   
## 1.285341 1.442055

## Interpreting the result

When reading the output of the Two-sample t-test the output follows a consistent format. The first line indicates the data used to calculate the test. These are the two vectors of data you listed in the t.test() command. The second line shows three values. t = -2.4484 which is the t-value calculated from the data, df = 27.756 is the degrees of freedom calculated for the data, p-value = 0.01047 is the p-value calculated.

The third line contains the alternative hypothesis. Notice that since we specified that the t-test is one sided, the alternative hypothesis is that the difference in means is less than 0.

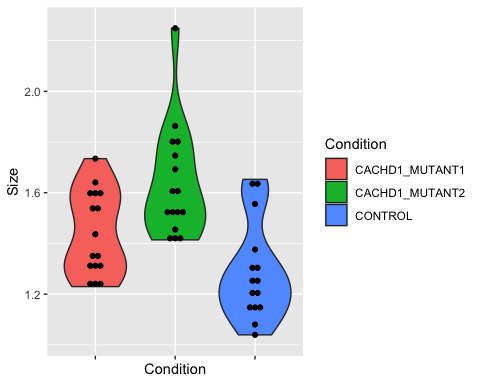
The fourth and fifth line contains the 95% confidence interval. The remaining information shows the means that were calculated for each population named x and y. R will assume that the first argument you give is vector x, and the second argument you give is vector y.

### Challenge 1A

Use a plot to depict the distribution of the three conditions at day 27.

ggplot(OrganoidSizeDay27, aes(x = Condition, y = Size,   
 fill = Condition)) +  
 geom\_violin() +  
 geom\_dotplot(binaxis = "y", stackdir = "center", dotsize = 0.5, fill = "black") +   
 theme(axis.text.x = element\_blank())

## Bin width defaults to 1/30 of the range of the data. Pick better value with  
## `binwidth`.



### Challenge 1B

It appears as though Mutant 2 may have an even more dramatic difference compared to the Control sample. Run a two sided t-test to compare if the average size of Mutant2 is different from the average size of the Control.

Mutant2Size <- OrganoidSizeDay27 %>%  
 filter(Condition == "CACHD1\_MUTANT2")  
  
t.test(ControlSize$Size, Mutant2Size$Size)

##   
## Welch Two Sample t-test  
##   
## data: ControlSize$Size and Mutant2Size$Size  
## t = -4.772, df = 28.837, p-value = 4.84e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.5017383 -0.2006350  
## sample estimates:  
## mean of x mean of y   
## 1.285341 1.636528

## Paired t-test

A paired t-test is sometimes considered a dependent sample t-test. This test is used when the samples may be dependent on one another. Let’s say there was a disagreement in the lab as to whether the CACHD1 mutant organoids reach their maximum size at day 18 or day 27. If we wanted to test to see if Mutant 1 grew significantly between day 18 and 27?

# Create a df for day 18  
OrganoidSizeDay18 <- OrganoidSize\_Mutant %>%  
 filter(Day == 18)  
  
# Create a df for the two days we will compare  
Mutant1\_Day18 <- OrganoidSizeDay18 %>%  
 filter(Condition == "CACHD1\_MUTANT1")  
Mutant1\_Day27 <- OrganoidSizeDay27 %>%  
 filter(Condition == "CACHD1\_MUTANT1")  
  
# Run the t-test, used `paired = T` to indicate it is a paired T-test.  
t.test(Mutant1\_Day18$Size, Mutant1\_Day27$Size, paired = T)

##   
## Paired t-test  
##   
## data: Mutant1\_Day18$Size and Mutant1\_Day27$Size  
## t = -0.87603, df = 15, p-value = 0.3948  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## -0.1304850 0.0544689  
## sample estimates:  
## mean difference   
## -0.03800803

The reason a paired t-test is appropriate in this scenario is beacuse the size at day 27 may depend on the size of the organoids at day 18.

### Challenge 2

We saw that Mutant 1 did not grow significantly between day 18 and day 27. Did mutant 2 grow significantly between day 18 and day 27?

Mutant2\_Day18 <- OrganoidSizeDay18 %>%  
 filter(Condition == "CACHD1\_MUTANT2")  
Mutant2\_Day27 <- OrganoidSizeDay27 %>%  
 filter(Condition == "CACHD1\_MUTANT2")  
  
  
t.test(Mutant2\_Day18$Size, Mutant2\_Day27$Size, paired = T)

##   
## Paired t-test  
##   
## data: Mutant2\_Day18$Size and Mutant2\_Day27$Size  
## t = -3.792, df = 15, p-value = 0.001772  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## -0.3866731 -0.1083974  
## sample estimates:  
## mean difference   
## -0.2475353

## Summary

T-tests are some of the most common statistical tests that you will see in biomedical research. These statistical tests use descriptive characteristics such as the mean and variance in order to compare between two samples. Different t-tests follow different assumptions based on the sample characteristics. In this section we covered Student’s t-tests for comparing samples with equal variance, Welch’s t-test for comparing samples with unequal variance, and paired T-tests for populations that may be dependent on one another.

# Analysis of Variance (ANOVA)

ANOVA statistical tests are used to analyze the difference between means of *more than two groups*. Next to t-tests ANOVA are also among the most common statistical tests used in biomedical research. Similar to t-tests ANOVA can be set up as one-way or two-way tests. The **null** hypothesis of ANOVA is that there is no difference among group means. The **alternative** hypothesis is that at least one group differs significantly from the overall mean of the dependent variable. ANOVA determines whether the groups created by the levels of the independent (qualitative) variable are statistically different by calculating the mean and variance for each group and comparing them against the overall mean of the dependent variable.

ANOVA tests follow a set of assumptions, which can inform when it should be used.

* Observations are independent from one another.
* The dependent variable is normally distributed.
* The variance between each group being compared is similar.

If your samples do not fit those assumptions the data may not be appropriate for using an ANOVA.

## One way ANOVA

A one-way ANOVA tests the differences in mean for one independent variable between more than two groups. You should use a one-way ANOVA when you have data about one categorical independent variable and one quantitative dependent variable. The qualitative variable should have at least three levels (representing the different groups to consider). We can use the OrganoidSize\_KD data to perform a one-way ANOVA.

Use the aov() command to run an ANOVA. The first argument contains the variables which you are using for comparison written using the following syntax: dependent ~ independent. The second argument is the data frame from which to draw the data using the following syntax data = data.frame. It is helpful to assign the output of aov() to an object. The summary() command is used to provide more descriptive information about the ANOVA output. See the example below.

OrgSize\_OneWay <- aov(Size ~ Condition, data = OrganoidSize\_KD)  
  
summary(OrgSize\_OneWay)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Condition 2 409.2 204.59 7.714 0.00206 \*\*  
## Residuals 29 769.2 26.52   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Interpreting the ANOVA results.

The ANOVA output creates an estimate for how much variation in the dependent variable can be explained by the independent variable.

* The first column lists the names of the independent variable(s): one for a one-way test, two for a two-way test and the model residuals.
* The Df column displays the number of degrees of freedom for the independent variable. For the dependent variable this is calculated by subtracting 1 from the number of levels found in the data. For the residuals it is calculated by subtracting 1 from the total number of observations minus the number of levels of the independent variable.
* The Sum Sq column displays the sum of squares, or the total variation between the group means and the overall mean explained by the variable.
* The Mean Sq column is the mean of the sum of squares which is calculated by dividing the sum of squares by the degrees of freedom.
* The F value is the test statistic (like the T-value from our t-tests). The larger the F value the more likely it is that the variance associated with the independent variable is not due to chance.
* Pr(>F) column is the p-value associated with the F statistic. On the far right side it will include significance codes indicating the level of significance show in the table below.

Based on the results from our ANOVA we can say that at least one of the conditions differs significantly from the group average. We see that we have an F value of 7.714 and a p-value of 0.00206, which indicates a fairly high level of significance. Importantly, an ANOVA test **does not** tell us which group is different, it simply tells us that there is a group that is different. We will see how to identify the different group (our groups) as we cover multiple comparisons.

## Two Way ANOVA

A two-way ANOVA tests the difference in one dependent quantitative variable for two independent variables with two or more levels each. Similar to a one-way ANOVA, a two-way ANOVA follows the same three assumptions listed above. In contrast, a two-way ANOVA tests three null hypotheses at the same time.

1. There is no difference in group means at any level of the first independent variable
2. There is no difference in group means at any level of the second independent variable
3. The effect of one independent variable does not depend on the effect of the other independent variable (there is no interaction between variables)

Like the one-way ANOVA, a two-way ANOVA calculates an F value my comparing the variance in each independent group to the overall variance of the dependent variable. If the variance within groups is smaller than the variance between groups, the F test will be higher and therefore less likely to be caused by random chance.

In our t-test example we reviewed data from knockout lines where the authors observed organoids grew to different sizes compared to control samples after a certain period. The authors measured the size of the organoids (our quantitative variable) for three conditions (the first qualitative independent variable) and eight time points (the second qualitative independent variable).

We can organize our hypotheses using the following framework:

| **Null Hypotheses** | **Alternative Hypotheses** |
| --- | --- |
| There is no difference in average organoid size for any sample condition | There is a difference in average organoid size by sample condition |
| There is no difference in organoid size for any day the sample was measured | There is a difference in average organoid size by day the sample was measured. |
| The effect of one independent variable on average organoid size does not depend on the effect of the other independent variable (there is no interaction) | There is an interaction effect between day the organoid was measured and the sample condition on average organoid size |

# organize your data frame to prepare for the anova test  
  
OrganoidSize\_Mutant\_clean <- OrganoidSize\_Mutant %>%  
 select(Day, Condition, Size, OrgID)

When handling two qualitative independent variables you can use a two-way ANOVA to test for additive effects, or for interaction effects. Notice the two examples below. In the additive two-way ANOVA we include the first argument Size ~ Condition + Day where we list the quantitative variable first, followed by the two qualitative variable added together.

OrgSize\_TwoWay\_additive <- aov(Size ~ Condition + Day, data = OrganoidSize\_Mutant\_clean)  
summary(OrgSize\_TwoWay\_additive)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Condition 2 3.79 1.89 60.22 <2e-16 \*\*\*  
## Day 1 57.39 57.39 1823.49 <2e-16 \*\*\*  
## Residuals 372 11.71 0.03   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

OrgSize\_TwoWay\_interaction <- aov(Size ~ Condition \* Day, data = OrganoidSize\_Mutant\_clean)  
summary(OrgSize\_TwoWay\_interaction)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Condition 2 3.79 1.89 67.22 < 2e-16 \*\*\*  
## Day 1 57.39 57.39 2035.64 < 2e-16 \*\*\*  
## Condition:Day 2 1.28 0.64 22.64 5.3e-10 \*\*\*  
## Residuals 370 10.43 0.03   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Interpreting the Two-Way ANOVA results

In the examples above we can see that both in our additive two-way ANOVA and our interaction two-way ANOVA, we have highly significant results. We see that there **is a significant difference** in organoid size based on **the day the sample was collected** AND based on **the experimental treatment condition**. Using the interaction test, it also appears that there is an interaction. These are results we might expect since we expect that the organoids will grow over time and that mutating a key gene (CACHD10) would change the behavior of growth. In this case the interaction is that as the days progress, the size changes in a nonrandom way that is different between the experimental conditions.

As is true for most things in statistics, determining which model is the best fit can be determined using a statistical test, in this case the Akaike information criterion (AIC) test. This test requires the package AICcmodav so let’s start by installing the package.

library(AICcmodavg)

The way the AIC test works is that it calculates the strength of each model by balancing the variation explained against the number of parameters used. The test will return a numeric value, which we can call AICc, indicating how much of the model **is not** explained by the statistical test. Therefore, the lower the AICc, the better the model to explain the data.

To run the AIC test, you use the aictab() command. This command requires two arguments, first a list of the models that you created, and a vector of corresponding model names in order to create its output. We can define these below:

model.set <- list(OrgSize\_TwoWay\_additive, OrgSize\_TwoWay\_interaction)  
model.names <- c("OrgSize\_TwoWay\_additive", "OrgSize\_TwoWay\_interaction")  
  
aictab(model.set, modnames = model.names)

##   
## Model selection based on AICc:  
##   
## K AICc Delta\_AICc AICcWt Cum.Wt LL  
## OrgSize\_TwoWay\_interaction 7 -266.56 0.00 1 1 140.43  
## OrgSize\_TwoWay\_additive 5 -227.29 39.27 0 1 118.73

The output presents several columns and it orders the models from the model that best explains the data to the one that least explains the data. Below we describe what each column means.

* K- the number of independent variables used to create the model.
* AICc - the AIC score: the lower the better.
* Delta\_AICc - the difference in AIC score between the best model and the comparison model listed.
* AICcWt - the AIC weight, or the proportion of total variation that the model explains, multiply this value by 100 to get the percent.
* Cum.wt - the sum of the AICc weights from each model
* LL - log-liklihood is the value describing how likely the model is, given the data.

The most important parts to pay attention to here are he Delta\_AICc and the AICcWt. In our example our OrgSize\_TwoWay\_interation model AICcWT is 1 and explains 100% of the cumulative model weight. If another model explains a substantial percentage of the model weight (more than 5%), check the Delta\_AICc. If the Delta\_AICc value for the second best model is more than 2 AIC units higher than the best model, it may not be significant enough to be reported.

Our result here is that our Two Way ANOVA test for interactions best explains the variance in our samples. The AIC also validates our expectations that both Day and Condiiton affect the size of the organoid in a significant way.

### Challenge 3A

Since we expect changes in organoid size based on the day the sample was measured, perhaps we don’t need to include the timepoint in our model. Run a one-way ANOVA on the OrganoidSize\_Mutant data set to see if the condition impacts organoid size.

Describe how you would interpret the results.

OrgSize\_OneWay <- aov(Size ~ Condition, data = OrganoidSize\_Mutant\_clean)  
summary(OrgSize\_OneWay)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Condition 2 3.79 1.8950 10.23 4.73e-05 \*\*\*  
## Residuals 373 69.09 0.1852   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### Challenge 3B

Re-run the AIC test comparing the three models we have created.

Describe how you would interpret the results.

OrgSize\_OneWay <- aov(Size ~ Condition, data = OrganoidSize\_Mutant\_clean)  
summary(OrgSize\_OneWay)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Condition 2 3.79 1.8950 10.23 4.73e-05 \*\*\*  
## Residuals 373 69.09 0.1852   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Dunnett’s Test for Multiple Comparisons

In the previous section we convered how to run one and two-way ANOVA tests. We also showed how we can compare different models using an AIC test in order to determine which model best explains the data. Keep in mind that ANOVA tests help you determine if one group is different from any other group in your sample, and if there might be interactions between variables. However, it does not tell you *which* group is different. You can follow up your ANOVA tests with Multiple Comparison tests such as a Dunnett’s Test or Tukey’s Test to determine which group is significantly different. The Dunnett’s method compares means from several experimental groups against a control group mean to see if there is a difference. In contrast, Tukey’s method would be used if there was not a control sample or group. In the case of our data, we can use Dunnett’s method since we have a designated control group. To run this test we will need to use the DescTools library.

library(DescTools)

To run the Dunnett’s Test, we will use the DunnettTest() command which will require three arguments: x =, g =, and control =.

* x = is where you define the quantitative dependent variable from your ANOVA test
* y = is where you define the qualitative independent variable from your ANOVA test
* control = is where you can include a character string that matches the way the control group is named in your data set.

See the example below for how we apply it to our organoid data set.

DunnettTest(x = OrganoidSize\_Mutant\_clean$Size, g=OrganoidSize\_Mutant\_clean$Condition, control = "CONTROL")

##   
## Dunnett's test for comparing several treatments with a control :   
## 95% family-wise confidence level  
##   
## $CONTROL  
## diff lwr.ci upr.ci pval   
## CACHD1\_MUTANT1-CONTROL 0.1810655 0.05973022 0.3024007 0.0020 \*\*   
## CACHD1\_MUTANT2-CONTROL 0.2378146 0.11647931 0.3591498 3.5e-05 \*\*\*  
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Interpreting the Dunnett’s test results

Four our Dunnett’s test output we get something similar to the ANOVA and the AIC output/summary: a table with several columns. This is how we might read it in context with our data set:

* The first column includes the names of the groups being compared, with the name of the control sample shown separated above the others.
* diff - shows the difference in size between the control group and experimental group
* lwr.ci is the lower limit to the 95% confidence interval for the sample
* upr.ci is the upper limit to the 95% confidence interval for the sample
* pval is the p-value calculated whe comparing the sample with the control

The the case of our results we can see Mutant1 has a p-value of 0.0020 and Mutant2 has a p-value of 3.5e-05. We can say with a high degree of confidence that both Mutant1 and Mutant2 are both significantly different from the control.

Notice that the Dunnett’s test looks at all the data points together, without separating by day. In the case of our research study, we might want to conduct a separate Dunnett’s test for each individual time point, rather than on the collective data set.

### Challenge 4

Use a Dunnett’s test to assess *which* timepoints are significantly different between the control and the experimental conditions.

Describe how you would interpret the results.

Day6 <- filter(OrganoidSize\_Mutant\_clean, Day == 6)   
  
DunnettTest(x = Day6$Size, g = Day6$Condition, control = "CONTROL")

##   
## Dunnett's test for comparing several treatments with a control :   
## 95% family-wise confidence level  
##   
## $CONTROL  
## diff lwr.ci upr.ci pval   
## CACHD1\_MUTANT1-CONTROL 0.03430625 -0.028097487 0.09670999 0.3544   
## CACHD1\_MUTANT2-CONTROL 0.05897132 -0.003432422 0.12137506 0.0661 .   
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Day9 <- filter(OrganoidSize\_Mutant\_clean, Day == 9)   
  
DunnettTest(x = Day9$Size, g = Day9$Condition, control = "CONTROL")

##   
## Dunnett's test for comparing several treatments with a control :   
## 95% family-wise confidence level  
##   
## $CONTROL  
## diff lwr.ci upr.ci pval   
## CACHD1\_MUTANT1-CONTROL 0.1496626 0.10455815 0.1947671 3.3e-09 \*\*\*  
## CACHD1\_MUTANT2-CONTROL 0.1367046 0.09160018 0.1818091 3.0e-08 \*\*\*  
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Day12 <- filter(OrganoidSize\_Mutant\_clean, Day == 12)   
  
DunnettTest(x = Day12$Size, g = Day12$Condition, control = "CONTROL")

##   
## Dunnett's test for comparing several treatments with a control :   
## 95% family-wise confidence level  
##   
## $CONTROL  
## diff lwr.ci upr.ci pval   
## CACHD1\_MUTANT1-CONTROL 0.1761545 0.10073039 0.2515787 6.3e-06 \*\*\*  
## CACHD1\_MUTANT2-CONTROL 0.1616020 0.08617786 0.2370261 2.7e-05 \*\*\*  
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Day18 <- filter(OrganoidSize\_Mutant\_clean, Day == 18)   
  
DunnettTest(x = Day18$Size, g = Day18$Condition, control = "CONTROL")

##   
## Dunnett's test for comparing several treatments with a control :   
## 95% family-wise confidence level  
##   
## $CONTROL  
## diff lwr.ci upr.ci pval   
## CACHD1\_MUTANT1-CONTROL 0.3532756 0.2154790 0.4910723 1.1e-06 \*\*\*  
## CACHD1\_MUTANT2-CONTROL 0.3382214 0.2004248 0.4760181 2.5e-06 \*\*\*  
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Day27 <- filter(OrganoidSize\_Mutant\_clean, Day == 27)   
  
DunnettTest(x = Day27$Size, g = Day27$Condition, control = "CONTROL")

##   
## Dunnett's test for comparing several treatments with a control :   
## 95% family-wise confidence level  
##   
## $CONTROL  
## diff lwr.ci upr.ci pval   
## CACHD1\_MUTANT1-CONTROL 0.1567136 -0.001324339 0.3147515 0.0522 .   
## CACHD1\_MUTANT2-CONTROL 0.3511866 0.193148699 0.5092245 1.5e-05 \*\*\*  
##   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Conclusion

In this section we have covered bivariate statistical tests including the widely used t-test. We described use cases for Student’s T and Welch’s T tests. We also described various tests for analysis of variance when you have multiple independent categorical variables to explain one quantitative variable including one-way ANOVA, two-way additive ANOVA and two-way interaction ANOVA. We described how to use an Akaike Information Criterion test if multiple ANOVA models reveal significant results. Finally since ANOVA tests don’t tell you *which* sample is different from the group, we described how to use a Dunnett’s test for multiple comparison to identify which samples differ significantly from a control sample.

As we continue, we will explore the statistical tests used in order to analyze and process increasingly complex data. We will introduce when to use linear modeling and how to interpret those results. We will describe multivariate statistics where we introduce principal component analysis and k means clustering to process multiple dependent output variables. Remember: descriptive statistics, statistical tests and data visualization are all used in conjunction with one another in order to make inferences about how sample data relate to the larger populations they aim to represent. As samples get more complex, statistical tools serve to simplify the complexity of the data in order to identify trends or patterns that lie within your data.

# Sources and Extra Resources

Source: [T-tests](https://www.scribbr.com/statistics/t-test/)  
Source: [Paired T-tests](https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/)  
Source: [One Way Anova](https://www.scribbr.com/statistics/one-way-anova/)  
Source: [Two Way Anova](https://www.scribbr.com/statistics/two-way-anova/)  
Source: [Running ANOVA in R](https://www.scribbr.com/statistics/anova-in-r/)  
Source: [Selecting Models with AIC](https://www.scribbr.com/statistics/akaike-information-criterion/)  
Source: [Dunnett’s Test](https://www.statisticshowto.com/dunnetts-test/)  
Sourge: [Interpreting Dunnett’s Test](https://www.statology.org/dunnetts-test-r/)