# Guide to Statistical Hypothesis Tests (F-test and t-test) for COMP 5660/6660

Sean N. Harris, Daniel R. Tauritz

September 3, 2021

## Review of Statistical Hypothesis Tests

After a series of experiments is completed, your results will include a collection of numerical data, for example a list of final fitness values produced by an evolutionary algorithm. However, these values do not represent the full set of potential outcomes of your experiment; they are just a statistical sample. Because of this, it is not enough to compare to experiment configurations by simply averaging your resulting fitness values and seeing which is higher, because those values are not fully representative. Statistical hypothesis tests are used to perform a more rigorous comparison between sets of values, and can tell you to within a given confidence level whether there is a meaningful difference between your experiments. For this course, we will use the **two-sample F-test for equality of variances** (hereafter referred to as an F-test, though there are other kinds), and the **two-sample unpaired t-test for equality of means** (hereafter referred to as a t-test, though there are other kinds). For each statistical comparison in your reports, you will be performing an F-test, followed by a t-test, in order to test whether the two populations have equal means, or if one experiment configuration produces significantly higher fitnesses than the other.

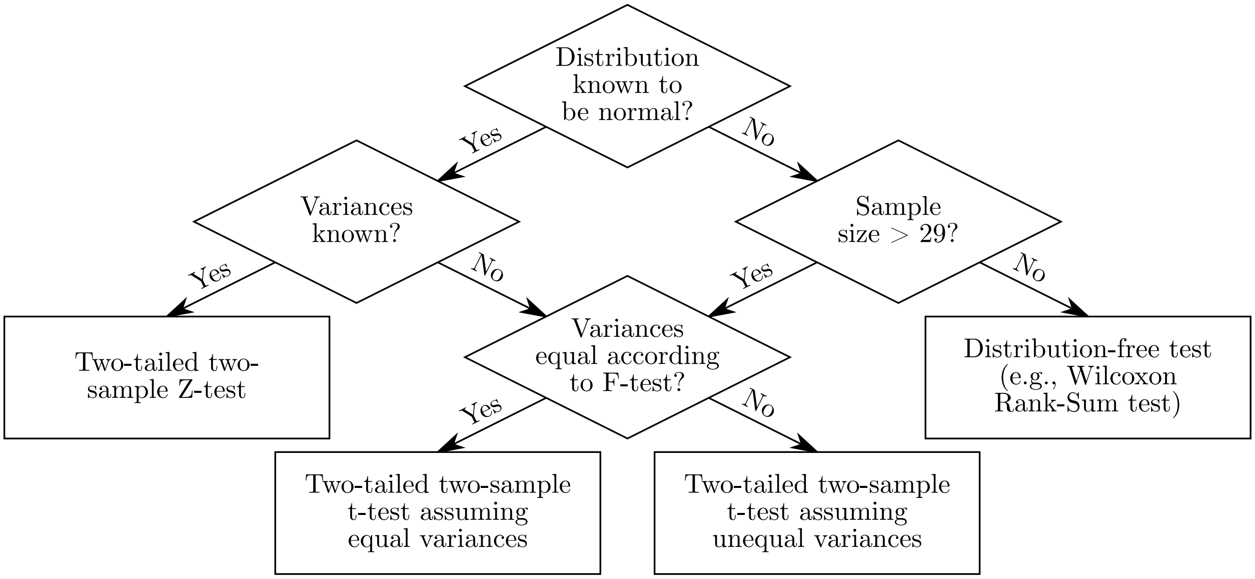
Distribution known to be normal?
Yes
No
Variances known?
Sample size > 29?
Yes
No
Yes
No
Two-tailed two-sample Z-test
Variances equal according to F-test?
Distribution-free test (e.g., Wilcoxon Rank-Sum test)
Yes
No
Two-tailed two-sample t-test assuming equal variances
Two-tailed two-sample t-test assuming unequal variances

Figure 1: Decision tree showing requirements for using a t-test

### Requirements for Use of the F- and t-test

Most statistical hypothesis tests work by assuming that your results follow a certain type of distribution, such as a normal distribution. This is good if your data is known to be sampled from a certain distribution, but this is generally not the case with experiment results. Instead, we follow a rule of thumb stating that taking an $n \geq 30$ sample size (30 runs of your EA configuration) tends to produce results that are close enough to a normal distribution to allow for the use of hypothesis tests that assume normality. **This is sufficient for your reports in this class**, however, in real-world usage there are more formal tests for normality of data which should be utilized, as sometimes your data won't produce a normal distribution even with a large number of samples.

### Significance Levels

When performing statistical tests, you must select a significance level ($\alpha$) to determine how highly sensitive you want the test to be, and equivalently how high a chance of type-I error you are willing to accept. In other words, when testing for a hypothesis that two values are distinct, $\alpha$ sets the probability that your test will incorrectly conclude that your hypothesis is true, when the two are not actually significantly different. **An $\alpha$ of 0.05 (5%) is commonly used as a default value for this, and should be used for experiments in this class**, though appropriate values differ by application and are often much lower.

## Technical Setup for Excel

These statistical tests can be performed using any tools that you are comfortable with, but we recommend using the Analysis ToolPak in Microsoft Excel. In order to run these tests automatically in Excel, the Analysis ToolPak must be enabled in the options. In Excel 2016 and later, navigate to "File → Options → Add-ins", click the "Manage: Excel Add-ins" option at the bottom of the menu, and check the "Analysis ToolPak" option in the window that appears. The Analysis ToolPak can then be accessed through the "Data → Data Analysis" ribbon section. Documentation for these functions is available at `https://support.microsoft.com/en-us/office/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6`

## Performing the F-test

The F-test for equality of variances tests whether the variances of two sampled populations are equal. This is important, because the t-test is sensitive to this fact, and will need to be conducted differently based on whether your two samples have equal or unequal variance. We will be performing a two-tailed F-test: we do not know ahead of time which if any population has the higher variance, as the population variances will differ from the sample variances, so we need to test simultaneously for positive and negative differences in variance with a two-tailed test. We use the F-test here for simplicity, however in practice more robust methods of comparing variance such as Levene's test are often preferable.

### Running the F-test Automatically in Excel

Excel only provides one-tailed F-tests, so in order to use it for two-tailed tests, you need to **halve the $\alpha$ value used from 0.05 to 0.025.** The process for running the F-test is as follows:

1. Arrange your data into two columns in Excel, one per experiment run.
2. Open the Data Analysis menu, and select "F-Test Two-Sample for Variances" option.
3. Set "Variable 1" to the first column, and "Variable 2" to the second column.
4. Set "Alpha" to half of your intended $\alpha$ value (0.025).
5. Set your desired output options and press "Ok".

Read the produced table and examine the entries labeled "F" and "F Critical one-tail". If F $>$ 1 and F $>$ F Critical one-tail, or F $<$ 1 and F $<$ F Critical one-tail, then the test has rejected the null hypothesis of equal variances, and determined that the two populations have significantly unequal variances. Otherwise, there is no significant difference in the two variances.

## Performing the t-test

The t-test tests whether the means of two sampled populations are equal. This is the main result that you're attempting to determine from your experiment data: whether one experiment configuration produces significantly higher fitnesses than the other. Based on the result of the F-test, we will use a t-test that assumes either equal or unequal variances. We will be performing a two-tailed t-test: we do not know ahead of time which if any population has the higher mean, as the population means will differ from the sample means, so we need to test simultaneously for positive and negative differences in mean with a two-tailed test.

### Running the t-test Automatically in Excel

For the t-test, by default **we will use an alpha of 0.05.** The process for running the t-test is as follows:

1. Arrange your data into two columns in Excel, one per experiment run. You can reuse the columns used for the F-test.
2. Open the Data Analysis menu, and select either the "t-test: Two-Sample Assuming Equal Variances" or the "t-test: Two-Sample Assuming Unequal Variances" option, depending on the result of your F-test.
3. Set "Variable 1" to the first column, and "Variable 2" to the second column.
4. Set "Hypothesized Mean Difference" to 0.
5. Set "Alpha" to your intended $\alpha$ value (0.05).
6. Set your desired output options and press "Ok".

Read the produced table and examine the entries labeled "t Stat" and "t Critical two-tail". If t Stat $>$ 0 and t Stat $>$ t Critical two-tail, or if t Stat $<$ 0 and t Stat $< -$t Critical two-tail, then your t-test found that the two experiments produced a significantly different mean fitness. The experiment with the highest sample mean can then be assumed to produce a significantly higher mean fitness than the other. Otherwise, we conclude the null hypothesis, and find that there was no significant difference detected between the two experiments.

## Required Output

**In your report, you need to provide at a minimum the following values from your F-test:**

- The chosen $\alpha$ value
- The sample size for samples 1 and 2 ("Observations" in Excel)
- The sample variances for samples 1 and 2 ("Variance" in Excel)
- The calculated test statistic F ("F" in Excel)
- Either the upper and lower critical values associated with the F-test, or the nearest critical value to F ("F Critical one-tail" for the latter in Excel)
- An interpretation of the outcome of the test and whether the result was significant

**Additionally, you need to provide at minimum the following values from your t-test:**

- The chosen $\alpha$ value
- Whether you performed your t-test assuming equal or unequal variances (the output header in Excel)
- The sample size for samples 1 and 2 ("Observations" in Excel)
- The sample means for samples 1 and 2 ("Mean" in Excel)
- The sample variances for samples 1 and 2 ("Variance" in Excel)
- The calculated test statistic t ("t Stat" in Excel)
- The upper and lower critical values associated with the t-test, or just the upper critical value ("t Critical two-tail" for the latter in Excel)
- An interpretation of the outcome of the test and whether the result was significant.

Except for the $\alpha$ values and the interpretations, these values will all be produced by Excel's F-test and t-test scripts in the tables it creates, and you can simply include the full tables in directly in your report. If you compute the F-test and t-test manually or through other software, you need to ensure that you provide all of these values in your report.

## Multiple Comparisons and Using Other Statistical Hypothesis Tests

In some assignments, you will be asked to compare more than two experiment configurations against each other. The easiest way to do this is to simply run pairwise t-tests for each pair of experiment configurations. **This is a sufficient methodology for this class.** However, in real-world practice, this can introduce unacceptable error, as each independent t-test adds further chance of error to the full analysis. If each t-test has a significance level of $\alpha = 0.05$, then a three-way comparison will have a familywise type-I error rate of 0.14, and a four-way comparison will have a familywise error rate as high as 0.46, nearly a 50% of incorrectly rejecting the null hypothesis! As a result, it is often preferable to either pick a smaller $\alpha$ to control for this (such as those given by a Bonferroni correction), or use alternative hypothesis testing methods such as ANOVA with Tukey's range test which more systematically control for error. The latter option can also be a less tedious way to run multiple comparisons, by performing them all in a single function. Additionally, perhaps the best way to limit error in your analysis is to simply not perform any unnecessary comparisons, and decide beforehand on the specific pairs of experiments that you are interested in comparing. If you are comfortable with these methods it is acceptable to use them, but otherwise **pairwise t-tests with an $\alpha$ of 0.05 will be considered sufficient for the purposes of these assignments.**