Guide to Statistical Hypothesis Tests for COMP 5660/6660

Sean N. Harris, Daniel R. Tauritz

August 25, 2023

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 **Welch's t-test**, which is an independent-samples t-test which does not assume equal population variances[1]. This will hereafter be referred to as a t-test, though there are other kinds. For each statistical comparison in your reports, you will be using the t-test as a location test in order to test whether the two statistical populations have equal means (the null hypothesis), or if one experiment configuration produces significantly higher fitnesses than the other.

Requirements for Use of Welch's t-test

Most statistical hypothesis tests work by assuming that your results follow a certain type of distribution; the t-test assumes that the sample means of your data are normally distributed. This is good if your data is known to be sampled from a normal distribution, but this is usually not the case with experiment results. However, the t-test is fairly robust to deviations from the normal distribution, particularly with large, similarly-sized samples, since the means approximate a normal distribution due to the central limit theorem. A typical rule of thumb uses an $n \geq 30$ sample size (30 runs of your EA configuration) to assume that the t-test is meaningful. **This is sufficient for your reports in this class.** However, in real-world usage the number of samples needed for the t-test depends on the characteristics of your data. EA results are frequently very non-normal even with large sample sizes, so it is usually better to use a non-parametric test such as the Wilcoxon-Mann-Whitney test[2], which has much more statistical power on non-normal distributions. A more detailed comparison of different tests and when to use them is given in the appendix.

Significance Levels

When performing statistical tests, you must first select a significance level (α) 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, α sets the probability that your test will incorrectly conclude that your hypothesis is true, when the two are not actually significantly

[1] As opposed to the better-known Student's t-test, which should only be used if it is known that the population variances are equal.

[2] Note that when you have differently-shaped distributions between your experiments, the Wilcoxon-Mann-Whitney test can only determine which experiment configuration was stochastically dominant, while the t-test can give confidence intervals for the effect size.

different. **An α 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

A wide variety of software tools exist to make statistical hypothesis testing convenient and easy. For this class, we recommend the SciPy library in Python, though this section also includes instructions for Excel for demonstration purposes.

Setup for Python

The SciPy library provides a large list of statistical tests under its `scipy.stats` module. The Conda environment provided for the class should come with SciPy installed. When working in another environment, ensure that your environment has the `scipy` package installed through PyPI. In your code, this module can be imported with `import scipy.stats`

Setup for Excel

Microsoft Excel comes with an add-in for statistical testing called the Analysis ToolPak. In order to use this functionality, 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 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. 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.

Once you perform your t-test, you'll receive a test statistic t representing how far away your observations are from the null hypothesis, and a p-value derived from t. This p-value represents the probability that you could receive a t at least this large if the null hypothesis is true. **If $p < \alpha$, then you can call the difference in sample means "statistically significant", and conclude that the experiment with the higher mean fitness performed significantly better.** Otherwise, we conclude the null hypothesis, and find that there was no significant difference detected between the two experiments.

Multiple Comparisons

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. However, 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 four-way comparison will have a familywise type-I error rate as high as 0.46, nearly a 50% probability of incorrectly rejecting the null hypothesis somewhere! As a result, it is often preferable to either pick a smaller α threshold per test to control for this, or use specialized hypothesis tests such as ANOVA with Tukey's range test which more systematically control for error. Additionally, perhaps the best way to limit error is to simply not perform any

unnecessary comparisons, and decide beforehand on the specific pairs of experiments that you are interested in comparing.

For assignments in this class with multiple comparisons, you should apply the Bonferroni correction, which is a particularly simple method of adjusting α to control familywise error. **When performing n different tests, each individual test compares its p-value against $\frac{\alpha}{n}$ instead of α.** This will ensure that the overall type-I error rate is less than α.

Running the t-test in Python

The function in SciPy for performing Welch's t-test is `scipy.stats.ttest_ind`, with the `equal_var` parameter set to False (otherwise it will incorrectly assume equal variances by default). An example function call is given below:

```
result = scipy.stats.ttest_ind(results_1, results_2, equal_var=False)
```
(1)

Where `results_1` and `results_2` are lists containing the best fitness value from each run. This function will return an object with several properties, including the t-statistic (`result.statistic`), p-value, (`result.pvalue`), and the number of degrees of freedom (`result.df`). From these, you can conclude whether your results had significantly different means. Documentation for this function is available at `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html`

Running the t-test in Excel

In Excel, the process for running the t-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 the "t-test: Two-Sample Assuming Unequal Variances" option.
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 α value (0.05).
6. Set your desired output options and press "Ok".

Your resulting table will list your t-statistic as "t Stat" and your p-value as "$P(T <= t)$ two-tail". Ignore the one-tailed values, as we are performing a two-tailed test.

Required Output

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

- The chosen α value (and the adjusted value if the Bonferroni correction is used)
- The sample size for samples 1 and 2
- The sample means for samples 1 and 2
- The sample standard deviations for samples 1 and 2
- The calculated p-value from the test
- An interpretation of the outcome of the test and whether the result was significant.

Appendix: Comparison of Hypothesis Tests

While the only statistical hypothesis test that will be used in this class is Welch's t-test, for completeness this section describes a few alternative tests, and when they are applicable.

- Student's t-test: Requires normally distributed sample means, equality of variance. Robust to moderate violations of these assumptions. Tests for a difference in means, and can provide a confidence interval.
- Welch's t-test: Requires normally distributed sample means. Robust to moderate violations of this assumption. Does not require equality of variance. Tests for a difference in means, and can provide a confidence interval. Nearly as powerful as Student's t-test, with fewer assumptions, so it is usually preferable.
- Wilcoxon-Mann-Whitney test: Requires no assumption about distribution in order to determine which population is "best" (stochastically dominant). For very non-normal distributions, this test has much more statistical power than a t-test! Can also be used to test for a difference in medians, if the two distributions are known to have the same shape. Otherwise, avoid this if you need to report an effect size.
- Median test: Requires no assumption about distribution. Tests for a difference in medians, even for differing distributions. Much weaker than the Wilcoxon-Mann-Whitney test, but can measure an effect size under broader conditions.