Continuous outcomes: Testing differences between groups

Overview:

If our outcome variable is continuous in nature, we often want to test whether there are any differences between groups of interest. For example, we may hypothesize that older individuals (65+ years old) have higher gene expression levels in a particular gene (our continuous outcome) compared to younger individuals (less than 65 years old). We assume that these groups are **independent** because each individual sampled is only in one of the two age groups and is assumed to be unrelated to anyone else in the sample.

An example of a comparison between **dependent** groups is if we examine differences in gene expression from a set of randomly sampled individuals at two different time points. Here, gene expression values from both time points are observed for each individual sampled. We now assume that measurements from the two time points are correlated because knowing the gene expression at Time 1 likely gives us some information about gene expression at Time 2 when measured from the same individual. Statistical tests that specifically account for this correlation are required with clustered data.

What statistical test do we use?

The form of the outcome variable (in this example, gene expression) dictates the statistical test you will use. As gene expression is continuous in nature, we will consider two main types of tests:

1) **Parametric:** T-test (2 groups), ANOVA (2 groups or more), Paired t-test (correlated data), linear regression
2) **Nonparametric:** Wilcoxon Mann-Whitney test (WMW, 2 groups), Kruskal-Wallis (2 groups or more), Wilcoxon Rank Sum (correlated data)

The main difference between these two types of tests is that parametric tests assume normally distributed expression levels whereas nonparametric tests do not. Nonparametric tests are rank-based, which is better when your data are skewed and/or have large outliers. **This document** provides some guidance on how to decide whether to use a parametric or non-parametric test for your data set. Also, **this document** provides R code for how to assess normality in your data.

Where can I read more about these tests?

There are many **excellent online resources** that will allow you to more fully understand how each statistical test is computed and the assumptions necessary to obtain valid results.

How do I perform these tests?

There are plenty of **statistical packages** and **online tools** that can be used to test for group differences. Here are some **basic instructions** for performing two-sample statistical tests (parametric and nonparametric) using the R software package.