Statistical Tests

You have learned hypothesis testing in the context, primarily, of t-tests for distributions. Using a t test you can check if a random sample implies that a population has a given mean, or if two populations have the same mean, or if one population mean has changed over a period of time. The framework you are using for hypothesis testing has its roots in procedures established by 19th-20th century mathematicians such as Karl Pearson, William Gosset, and Ronald Fisher and later by Jerzy Neyman and Egon Pearson. The process of formulating a null hypothesis which is to be disproven by overwhelming statistical evidence in favor of an alternative hypothesis is due to the pioneering work of these statisticians and underlies much scientific research published today.

However there are a vast number of other statistical distributions and tests that can be used in the framework of hypothesis testing. We will discuss some of those distributions and tests here.

The t-test is a very useful test but it has a number of prerequisites that limit its validity and applicability. In particular, t-tests require that the distribution of sample means is normally distributed. This is guaranteed if the underlying population distribution is known to be normal, and can to a lesser degree, be assumed if the number of samples taken from the population is "large" (like 30 or more). In practice, establishing that a t test is valid requires the analyst to look at the data and make some fairly subjective claims about the data being non-skewed, and almost normal, and having few outliers, and therefore asserting that the t test is valid. It is then up to the reader to decide independently if the analyst's claims are sufficiently well justified. It is all, unfortunately, a rather large gray area.

As an alternative to t-tests there are a number of *non-parametric* hypothesis tests that do *not* require any assumptions about the normality of the data, or the presence of outliers, or even the shape of the distribution. These tests are all much easier to justify in practice and, therefore, deliver results that could be more plausible to the scientific community.

So why not always use non-parametric tests? They come with one weakness – they are not as discriminating as a parametric test. That is, it's harder, in general for a non-parametric test to reject the null hypothesis. This means, as a researcher, it takes more unexpected data to get a publishable result. In statistical terms, we say the non-parametric tests are less "powerful" than t-tests because the probability of rejecting H_0 when H_a is true is lower for a non-parametric test. On the other hand, the result itself, by relying on fewer assumptions, makes a stronger argument than a t-test when a result is found. As one author puts it, non-parametric tests are "always valid, but not always efficient," while parametric methods are "always efficient, but not always valid."[1]

[1] Nahm F. S. (2016). Nonparametric statistical tests for the continuous data: the basic concept and the practical use. Korean journal of anesthesiology, 69(1), 8–14. https://doi.org/10.4097/kjae.2016.69.1.8

Review of T Tests

For a one-sample T test, given an assumed mean μ_0 the analyst takes a random sample and computes a sample mean $\bar{X}$ and sample standard deviation S_X. The sample is then compared to the mean using $T = \dfrac{\mu_0 - \bar{X}}{S_X/\sqrt{n}}$. The more $\bar{X}$ deviates from μ_0, the larger this value will be, and the more likely rejecting the null hypothesis becomes. This test (and the two sample and paired sample versions) rely on several *parameters* of the population distribution. They assume we know, or can estimate, μ and σ, and that the accuracy of our estimates can be bounded because we know the sample means $\bar{X}$, are normally distributed.

Intro to Non-parametric Tests

Non-parametric tests do not rely on any assumptions about the distribution of the sample or the population, or on parameters such as the mean or standard deviation of the population. Instead of comparing sampled values to the sample mean, elements in a sample are typically sorted and calculations are performed on the *rank* of the data. The actual values, once sorted, are discarded. This allows non-parametric tests to draw conclusions about the general shape of a distribution, without needing to approximate the mean or standard deviation.

Non-parametric test hypothesis are usually stated as "X and Y are samples from the same distribution" or, sometimes, "X and Y are drawn from distributions with the same median" or, sometimes "the same location." (I've been told that AOS science classes tend to rely on "median" for hypothesis tests so that saves us quite a bit of trouble.) To see how complex the correct null hypothesis can be, you can peruse this wikipedia article[2].

One-Sample Wilcoxon Signed-Rank Test

The first test we will overview is a non-parametric version of the one-sample t test. This test (roughly) compares a sample of data to an assumed median. It works (roughly) by sorting the data and labeling points above the median as positive and those below the median as negative. The ranks of the values (not the values but their **ranks**) are combined to give a statistic such that a large statistic implies agreement with the null hypothesis.

Example 1: One sample test

Here's a quick example of data with supposed median of 0. After sorting the data, it is ranked, where the rank is *signed* (the sign of the rank equals the sign of the data).

[2] https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

data	**abs(data)**	**rank**
-32	32	-7
-21.4	21.4	-6
-10.2	10.2	-5
-3.2	3.2	-2
2.6	2.6	1
3.5	3.5	3
4.1	4.1	4

Now we look at the sum of the positive ranks: $1 + 3 + 4 = 8$ and the sum of the negative ranks: $7 + 6 + 5 + 2 = 20$. If the data were perfectly symmetric around 0, these two sum would be the same at 14 each. The smaller one, 8, is our test statistic. Using a computer or a table for a 2-sided Wilcoxon Signed-Rank test gives a p-value of 0.375. There is no reason here to reject the null hypothesis that the median is 0. By comparison, the t-test for this data give a p-value of 0.177, and the one-sided t-test p-value is 0.085, which is getting very close to 0.05. Picking the appropriate test can have a big impact on your result, and its validity!

Example 2: One sample test

In this example we'll take a set of data: 87.5, 92.1, 92.3, 100.2, 101.7, 103.5, 105.7, 109.2, 112.5, 121.1 and compare the median to an assumed value of 93. For this test, we use the same strategy, but subtract 93 from each data point first:

x	$x - 93$	$abs(x)$	rank
87.5	-5.5	5.5	-3
92.1	-0.9	0.9	-2
92.3	-0.7	0.7	-1
100.2	7.2	7.2	4
101.7	8.7	8.7	5
103.5	10.5	10.5	6
105.7	12.7	12.7	7
109.2	16.2	16.2	8
112.5	19.5	19.5	9
121.1	28.1	28.1	10

The signed rank statistic here is $1 + 2 + 3 = 6$ is quite small and yields a p-value of 0.027, suggesting the median does differ from 93.

Example 3: Paired test

The Wilcoxon signed rank test can also be used instead of a *paired t-test* when a non-parametric approach is more appropriate. Recall a paired t-test is used to compare one sample to itself at 2 points in time, such as when a test group is given a drug to see if some symptom improves. In this scenario, observations are in the form (x_1, x_2) – one pair for each point in the sample. If there is no difference over the two points in time, you would expect the median of $x_2 - x_1$ to be 0. And this is exactly what we test.

Consider a class whose quiz scores are taken before and after attending a tutoring session. The grades before and after for each student are given in the table below, along with a Paired Wilcoxon Signed Rank Test.

x_1	x_2	$x_2 - x_1$	$\|x_2 - x_1\|$	**rank**
17.5	19.1	1.6	1.1	5
16.5	17.1	0.6	0.6	2
18.5	18.0	-0.5	0.5	-1
16.0	19.5	3.5	3.5	8
19.0	17.5	-1.5	1.5	-4
20.0	21.0	1.0	1.0	3
13.5	16.5	3.0	3.0	7
12.0	14.0	2.0	2.0	6

The statistic here is somewhat small $1 + 4 = 5$ but the p-value of 0.10 is not significant.

Two sample non-parametric test: Mann Whitney U Test

The Wilcoxon test works for both one sample and paired sample cases. When there are two independent samples, a different approach is needed. This one we will just overview without getting into the details, but it is also based on sorting and ranks. Let's say you have a sample of size 5 of values from one population and a sample of 7 from another. If you sorted all 12 values and then replaced the numbers with and X or an O depending on which distribution they came from you could get something like the following

- 'XXOOOXOOOXOX'
- 'XOXOOOOXXOXO'
- 'OOOXXXOOOXXO'
- 'OOOOOOXXOXXX'
- 'XXOXXOXOOOOO'

The first 3 of these look fairly balanced in terms of where the X and O are. But note the last 2. X is generally larger in one, and generally smaller in the other. Just by looking at the overall rank of each sample, we can roughly determine if the distributions are close or quite different. The Mann-Whitney U test makes this precise and handles null hypothesis such as "X and O come from the same distribution" or (more likely in AOS) "X and O have the same median."

We won't include a sample here but you should try running one on Datascape just to get a feel for the test.

Other tests not covered yet

There are a few other tests to be familiar with, listed below. You can do all of these in Datascape.

- If you need to compare 3 or more distribution means: ANOVA
- A non-parametric version of ANOVA: Kruskall-Wallace
- If you want to compare data to a distribution (are birth months uniformly distributed? Are test scores normally distributed? Are these groups distributed similarly?): Use Chi Squared Goodness of Fit
- If you want to see if two factors are independent (e.g. gender vs. political party) (e.g. zip code vs. car model for seniors), use Chi Square Independece

In Conclusion

Parametric and non-parametric tests each have their place. The right test to use is often a subtle question involving thorough data analysis that we won't really get into here. But here are some good pointers

Parametric (t-test)

- Assume the population is normal, or sample size is large, or sample is almost normal
- Sometimes assumes variances of samples are the same, or close to the same
- Is more efficient, or more powerful, which means a smaller sample size is needed to prove an effect is real (compared to non-parametrics)
- Can be thrown off by outliers or skewed data

Non-parametric

- No assumptions about distribution or mean or variance
- Much easier to justify
- Works with small sample size (although parametrics can as well if normality is assumed!!)
- Less powerful, less efficient
- Robust to outliers
- Null hypothesis can be subtle (median, vs. location vs. distribution)
- Sometimes assumes symmetry of distribution about the median (we didn't really get into this but it's a non-trivial part of this topic)
- Can get tricky when you have "ties" or "zeros" in your data. Computers can take care of it, but you need to be aware of what's going on (we also didn't get into this)

My impression seems to be that all the AOS science experiments fall back on non-parametric tests just because it's easier to justify with small sample sizes. That's fine and makes your life easier. That said, knowing when parametric tests are valid gives you more leeway, and a better chance of getting a result!

Homework

Analyze the following datasets on Datascape. Pick an appropriate parametric or non-parametric test. State which test you used and why (include discussion of data graphical analysis) and discuss your conclusion.

- batteryLife2 compare A vs B and determine if one lasts longer5
- medicaidTimes compare the wait times to 18 minutes and see if there is a difference
- beetFertilizer compare the results of newA vs newB
- headHeight pick 2 classes where you can demonstrate a difference in height (the units here are height in cm divided by size of head in cm)