**Hypothesis Tests between Groups**

Now we’d like to consider tests for hypotheses about differences in groups.

**Null Hypothesis**

Our Null Hypothesis will often be, ‘there is no difference between the groups’. Most of the time this is insufficient to formulate a hypothesis. So we’ll require some additional information, like, the hypothetical means and standard deviations of the population. Often we’ll assume the population mean is the average of the two sample means. And

Apropos binomial proportions, we’ll often presume p0 = 0.50, or otherwise equals the average of the two proportions. Apropos multinomial proportions, looks like we generally have some other alternative hypothesis in mind.

**Alternative Hypothesis**

This would be that the groups are as different as they, or more.

**The tests**

We will often formulate a random variable equal to the difference in means between the two groups, or perhaps the difference in standard deviations if we’re interested in that. With two different populations, we will have two different sample standard deviations probably. If we can presume they actually have the same standard deviation, then we can use a pooled T-statistic. If not, I think we can use the Welch’s T-statistic. If we’re testing multiple groups, then we’ll use Multiple T-tests, and have to be cognizant of the necessary corrections. If we testing for differences in proportions between groups, binomial or multinomial, then one thing we can do is the χ2 test for independence, which is a modification of the χ2 test in the last file.

**Multiple/Pairwise T-tests for differences in means, and the Bonferroni correction**

Often we have a situation where we’re comparing means between groups. An example would predicting the average height of different age groups, or weight lost in different dieting plans. The Null hypothesis will often take the form of the assumption that there is no difference in average heights or average weight lost, etc. And the Alternative hypothesis will be that there is. When we have multiple values for a category, like if we split the age groups into 8 ranges, or if we’re testing 6 different diets, then we have to test each pair for whether there is a statistically significant difference in their mean outcomes, usually using a T-test. When have such a decent number of groups, the odds of making a false-positive (i.e., rejecting the Null Hypothesis and accepting the Alternative Hypothesis) increases. This is because in a large group, the odds that at least two of the groups will have picked up relative outliers and caused them to have an abnormally large difference between them increases. For instance, if we had 10 groups, then we’d be making 10∙9/2 = 45 comparisons. And even if we’d expect at the 95% confidence level that none of the groups differ, still, the odds that at least one out of the 45 groups will exceed the 5% p-value threshold is 1-(0.95)45 = 0.91. So very high. Thus, we need to make a ‘correction’ to the p-values to account for this. One possible, and popular, correction is the *Bonferroni correction*. I don’t know how to implement this myself, but in the pingouin package we can implement it.

**Rank-Sum Test (Nonparametric) for difference in means**

Non-parametric tests are useful because they do not presume the underlying population has a normal distribution. And they tend to work better for small sample sizes. So here’s the Rank-Sum Test. Say we have the following data, and we suspect Group 2 has a different mean value, perhaps higher, than Group 1 does.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

So we rank the numbers in order from smallest to largest. And if two numbers share the same rank, then we assign to them both the average of the two ranks they would’ve otherwise had if they weren’t the same value.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 4 | 7 |
| 6 | 2.5 |
| 2.5 | 8 |
| 1 | 5 |

Then we add up the ranks.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 13.5 | 22.5 |

Now our Null Hypothesis is that these groups have the same mean, and so all of these ranks could really have appeared in either column, and any row. So the probability that we’d get a Group 2, say, rank sum x that we got is:



where n is the number of rows. This comes from the fact that each arrangement of ranks for a particular group has a multiplicity of n!. We have n! ways to arrange the ranks in group 2 with out affecting the sum, x. And we have n! ways of arranging the ranks in group 1, obviously w/o affecting the rank in group 2. And there are (2n)! ways to assign any rank to any group. The p-value for getting a sum as large or greater than x\* would be:



Similarly, the probability of getting a sum smaller than x\* would be:



And the probability of getting a sum more extreme is:



where I’m calling μ the median rank sum, out of all possible. μ would be kind of hard to figure out. So normally, one backs their way into such a calculation by tabulating the sums going from most extreme to less extreme. Book says this is equivalent to the Mann-Whitney Test.

**ANOVA for differences in means**

There’s another way to test for differences means between groups. This method compares the variation of each group’s data about its own mean (SSEWG). And compares that to the variation of the group means themselves (SSEBG). If the former is much smaller than the latter, then that suggests the means are indeed different.

For visualization’s sake, consider picture with the means displayed, for three groups.

Chart, box and whisker chart

Description automatically generated

and we may roughly visualize the three SSE’s as (taking top red datapoint for example):

Chart

Description automatically generated

To proceed, we’ll first show that we can break down the total sum of square errors about the overall mean, SSEm, into the sum of two parts, SSEWG and SSEBG.



where we clearly define the squared error within groups, and squared error between groups as:



and have that:



At least under the assumption that all groups have the same mean and variance (our Null Hypothesis), it turns out SSEBG is a χ2 distributed variable with νBG = G-1 d.o.f., and SSEWG is a χ2 variable with νWG = n-G d.o.f., where n is the total number of data points. Now recall from the Single Variable PDF file that the ratio of two independent χ2 variables, divided by their d.o.f.,



is an F-distributed variable, with ν1 and ν2 d.o.f. So we can say,



So we have, where x = F,



and we’ll observe that if SSEWG << SSEBG, then F will be large, which will make the p-value:

p = ∫F\*∞ pF(x,d1,d2)dx small, in agreement with the general argument that if variation about the means is smaller than variation of the means about the overall mean, then the means are likely distinct, in contradiction to the Null Hypothesis. In practice, one often combines ANOVA tests and multiple T-tests. We can use ANOVA to verify that at least one of the means in the groups are different. And then we can do multiple T-tests to see which ones are different.

**Kruskal-Wallis Test (Nonparametric) for difference in means**

This guy is similar to the ANOVA test, but makes no assumptions about the underlying population statistics. It works like this. Say we have n total data points, which we have separated into groups indexed by i = 1, 2, .., G, each with ni elements. Then we collate all the data points, and rank them. Then we find I the average rank for each group, and compare to , the total average rank. We compare these two by forming the quantity analogous to SSEBG in ANOVA.



In this case we form,



We’ll note that is just the sum of 1, 2, 3, 4, 5, …, n, divided by n, which is n(n+1)/2n = (n+1)/2. So we can write,



Then they formed the statistic,



And this statistic follows what I’ll call the Kruskal-Wallis distribution with ν d.o.f. Letting x = H, we have, formally,



Turns out if the number of data points is large (and subset data points is large enough too?), then this distribution is approximately χ2,



If our test statistic x = H\* is large, then that would support rejection of the Null Hypothesis that the groups’ means do not differ. So we’d calculate a p-value,



Note this tests for whether there is a statistically significant difference in the means, not for the direction of the difference.

***Matched Pair* T-test for difference in means**

There’s a slightly different scenario where we’d like to test for a difference in means. This is when we have matched pairs. This case might arise if we apply a test to a group of students before lunch and then another test after lunch. And we’d like to see if having eaten lunch makes any difference in the scores. This kind of test might be attractive because we eliminate the possibility of confounding variables? But a problem with the test is that we cannot expect the values in Group 1, say X1 and the values in Group 2, say X2, to be uncorrelated. For instance, whatever X1 is, we would likely expect X2 to be higher. So they are not independent variables.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

However, I think the variable X1 – X2 does constitute a set of independent random variables? Not sure, considering I just said we can expect X2 > X1 or something. Either way, to look for a difference in the means, we consider the difference.



The mean of will just be the difference of the two means. But since the Xki’s are correlated, the variance of is not the sum of the variances – see Probability distribution file. Nonetheless, we could work out the variance, and form the variable,



and I think this would be normally distributed. But typically, we are not privy to the population variance, though we could approximate it with . So we form the T-variable instead.



where,



and ΔXj = X1j – X2j. And this is T-distributed with n-1 d.o.f. (n = number of data points). So letting,



we have:



So to calculate whether Group 1 is less than Group 2, say, given the Null Hypothesis, we’d form our measurement statistic (note we employ Null Hypothesis assumption μ1 = μ2),



And calculate,



and likewise for other alternative hypotheses.

***Matched Pairs* Sign-Test (Nonparametric) for difference in means**

The previous test presumed we knew something about the underlying population statistics when comparing our groups. We can disavow any such knowledge and still perform tests. These won’t be as precise as the preceding, but are more general. So say we have a bunch of data from two groups.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

We don’t presume to know anything about underlying statistics of the distributions of numbers in Group 1 or Group 2. But we’d like to know if Group 2, say, has a higher mean value than Group 1. One way to proceed is to take the differences of the groups,

|  |
| --- |
| Group 2 – Group 1 |
| 5 |
| -5 |
| 9 |
| 3 |

And we’d reason that if the groups had the same mean, then there should be as many positive differences as negative differences. Or better put, the probability of any row being positive is p = 0.50. This is our Null Hypothesis. So the probability of having x positives out of n rows is:



which simplifies to just,



And then we’d calculate the probability of having as many or more positives as we do, given our Null Hypothesis. If this works out to be less than say p = 0.05, then we accept the Null Hypothesis (provisionally). So our p-value is:



and we’d calculate similarly left-tailed



or two-tailed tests.



***Matched Pairs* Wilcoxon Signed-Rank Test (Nonparametric) for difference in means**

The previous method did not take into account the magnitude of the differences, just the sign. This method takes into account the actual numbers. So say we have a bunch of data from two groups.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

We don’t presume to know anything about underlying statistics of the distributions of numbers in Group 1 or Group 2. But we’d like to know if Group 2, say, has a higher mean value than Group 1. Again, we take the differences of the groups,

|  |
| --- |
| Group 2 – Group 1 |
| 5 |
| -5 |
| 9 |
| 3 |

Then we rank the absolute values of the differences. And if two differences tie, then we give them the average of the two ranks they would’ve otherwise had.

|  |
| --- |
| Group 2 – Group 1 Ranks |
| 5 (2.5) |
| -5 (2.5) |
| 9 (4) |
| 3 (1) |

Then we calculate the rank sum for positive and negative differences separately,

|  |  |
| --- | --- |
| Group 2 – Group 1 Ranks Sums | |
| T+ = 7.5 | T- = 2.5 |

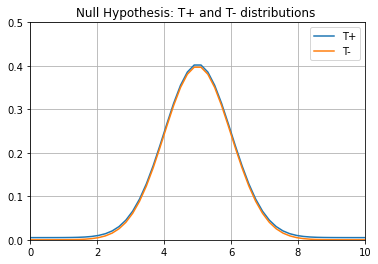
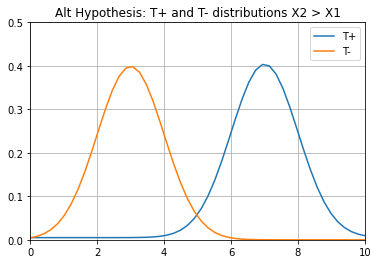
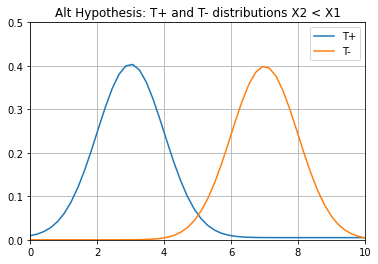
These two guys follow a probability distribution, which I’ll call the Wilcoxon pdf, with I’ll say ν = n d.o.f., where n is the number of rows. Letting x = T+, or T-, we can say:



As one might expect, for large n > 25 or so, this distribution is approximately normal,



since x is basically the sum of random variables and such a thing approaches a Gaussian distribution by the Central Limit Theorem. Can also see that the average and variance of pW(x;n) grows with n, as would be expected. Under the Null Hypothesis, the two distributions should coincide, as shown in the left graph. If Group2 values are greater than Group 1, or less than Group 1, then the T+, T- distributions should shift as displayed.

So we can use either T+ or T- as a test statistic to get a p-value for the likelihood of a measurement of T+ or T-, given the Null Hypothesis. Seems we typically use T-? If our Alternative Hypothesis is X2 > X1, then T+ ought to be large-ish, and T- ought to be small-ish. So we can test for abnormally small T-, or abnormally large T+ values. So we calculate the probability, according to the Null Hypothesis, that T- is less than our measurement T-\*, or T+ is greater than our measurement T+\*.



If our Alternative Hypothesis is X2 < X1, then T+ ought to be small-ish, and T- ought to be large-ish. So we can test for abnormally small T+ values, or abnormally large T- values. And we calculate the probability, according to the Null Hypothesis, that T+ is less than our measurement T+\*, or T- is larger than our measurement T-\*.



If we’re simply testing for a difference in the two, |X2 – X1| > 0, then we pick the min(T+\*,T-\*), and calculate



**χ2 Test for differences in proportions**

Another common scenario for consideration is when have two different experiments, each with two or more outcomes. And we might try to estimate whether the difference in outcomes of the experiments can be ascribed to some real underlying difference in the two experiments, or just due to random chance. For example, let’s say we have two experiments: 1,2, with three outcomes each: a, b, c.

|  |  |  |  |
| --- | --- | --- | --- |
|  | a | b | c |
| 1 | n1a | n1b | n1c |
| 2 | n2a | n2b | n2c |

The njℓ’s tabulate the outcomes for each experiment, j. And our question is whether the n1ℓ differ from the n2ℓ by chance or not. Or more to the point, whether the proportions n1ℓ/n1 differ from n2ℓ/n2 by chance or not (ℓ = a, b, c). Or put another way, whether P(ℓ|1) signficantly differs from P(ℓ|2) or not. If our Null Hypothesis is that they do not differ, then we’re saying that:



which implies,



So in effect, we’re saying that the ℓ events and j events are independent from each other. What remains is to estimate the probabilities P(ℓ) and P(j). We’d say,



Moreover then,



Which is:



Another way to look at this is, we get njℓ(pred) by locating element jℓ, summing the row and column its in, multiplying the two sums together, and dividing by n. Then define the variable Z2:



This will be χ2 distributed with ν = (ℓmax – 1)(jmax – 1) d.o.f. This is because we have this many d.o.f. in choosing the probabilities of the entries in the table, because we presume independence, so we have ℓmax and jmax probabilities to independently specify, but then the probabilities of each event must sum to 1, so that leaves us with (ℓmax – 1)(jmax – 1) d.o.f.



And this can be straightforwardly generalized to the case of more experiments than two, and more outcomes than three.

**Example: Pair T-test**

Say we take 50 people, and put 20 on a low carb diet, and 30 on a low fat diet. At end of 4 weeks, the low carb diet people lose average of 6lbs with (sample) std of 2 lbs. And the low fat people lose average of 5lbs with (sample) std of 3lbs. Question is, can we conclude, at 5% uncertainty level, that there is a meaningful difference in outcomes?

H0: Let group 1 be the low carb diet, and group 2 be the low fat diet. Null Hypothesis is that there is no difference, and that η0 = μ1 = μ2. And we’ll be interested in calculating the probability distribution of the difference, m, in average weight losses between the two groups. Like with our example above, just knowing that the average of the differences is zero (according to the Null hypothesis) isn’t enough to know Pη0(m). We’ll have to presume something about the variances too. We’ll get to that.

HA: The Alternative Hypothesis is that there is a meaningful difference, and μ1 ≠ μ2.

Test: So we’d like to work out the probability distribution for the difference of the two sample means, presuming the Null hypothesis that the two distributions the samples are pulled from should have the same mean. So let the measurement random variable be:



1,2 ought to be approximately normally distributed. And so M will be as well. The mean and variance of M are:



Or, putting in terms of the Z-score,



ought to be unit normally distributed. But yeah, we don’t know the population variances. What we can do is approximate σ1,2 with s1,2, the sample standard deviations.



where s12 = 22 = 4, and s22 = 32 = 9, and s12/n1 + s22/n2 = 4/20 + 9/30 = 0.50 ≈ 0.72. But since one of our samples is less than 20, it seems better to use the more accurate Student’s T distribution. So we would form,



where,



And then this variable is Student’s T distributed with number of d.o.f. equal to:



(from Welch’s T-test distribution – see Multiple Variables PDF). But this guy says it’s better to approximate it as ν = min(n1, n2) – 1. Whatever. Going with the guy, then we’d have ν = 20-1 = 19. And so,



The variance of this distribution is σ2 = ν/(ν-2) = 19/17 = 1.1 → σ = 1.05. But as this is pretty close to our standard normal distribution, I’ll just use that. Our measurement was:



So we want to know:



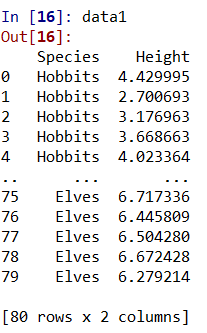
This is:



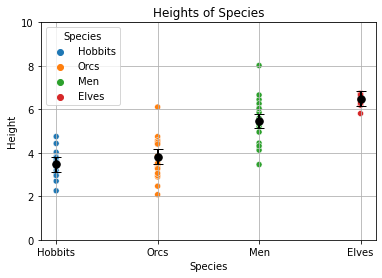
So we would accept the Null Hypothesis, and say that such a difference is small enough that their actual means could still be equal.

**Example: Multiple Pair T-test**

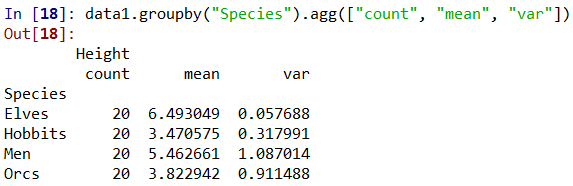
Say we have multiple groups that we’re comparing means for. We could run T-tests for all of them. If we have n groups, then there would be n(n-1)/2 pairs to compare. I made a data table of heights for different species in middle earth.



And did scatterplot, and overlayed error bars representing a 95% confidence interval for their averages.



Can see the species are fairly distinguishable, except for hobbits and orcs. Let’s do T-tests to compare the species and see if their heights are statistically significantly different. We’ll start with Hobbits and Orcs. Well, I made a datatable of means and (sample) variances for the different groups.



To compare the difference between Hobbits and Orcs, we’ll use the pooled T-test (see Multiple Variables PDF’s). For Hobbits and Orcs, the pooled variance is:



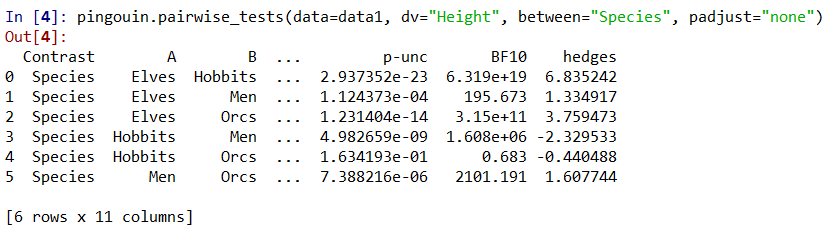
Then our test statistic is (presuming Null Hypothesis that there is no difference between the means):



So then, working out the area under the Student’s T distribution for |t| > 1.42, we have a p-value of p = 0.164. Thus, at the commonly accepted 95% confidence level (p = 0.05), we would fail to reject the Null Hypothesis that Hobbits and Orcs have the same mean height.

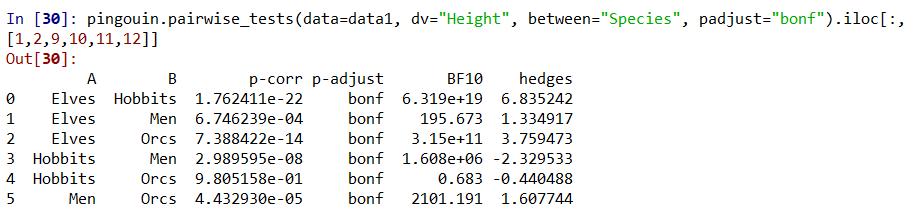


Note we used T distribution with n1 + n2 – 2 = 20 + 20 – 2 = 38 d.o.f., same as the d.o.f. in Sp. And we could keep comparing groups. We can do this with automatically with *pingouin*,



And we’ll observe our calculated p-value for Hobbits-Orcs compares well with its result. BTW, since all the samples have the same counts (20), pingouin will automatically use the pooled T-test, just as we did ourselves. But we could have used the Welch’s T-test, which it would automatically have done had the sample counts been different between groups.

We can implement this in pingouin by changing the padjust value from “none” to “bonf”. And we get (just isolating the relevant columns),



Can see that our p-values generally went up, making it less likely that we’d reject the Null Hypothesis. In particular the Hobbit-Orcs p-value went from 0.164 to 0.98!

**Example: ANOVA**

Let’s do the example above, but use the ANOVA test this time. So our measurement random variable is:



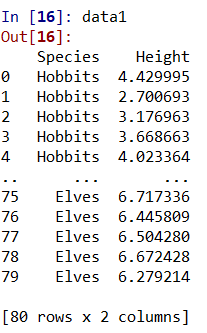
which will be F-distributed with 3 and 76 d.o.f. in numerator and denominator respectively. Calculating these guys,



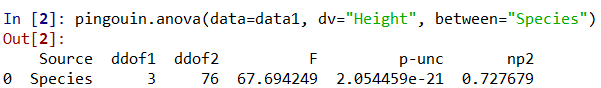
And then we calculate,



So our ANOVA test clearly indicates that *some* (at least one) of the means are different. We can do the ANOVA test directly from pingouin. The data must be in the form,



and then we say,



So we see the same F-statistic value. But apparently there’s a mismatch between calculating the integral in scipy vs. pingouin, as they get a p-value of 2.05×10-21.

**Example: Pair T-test**

Consider stats on two treatment options for proximal humerous fractures. We’d like to know if there is a statistically significant difference in outcomes.

A picture containing graphical user interface

Description automatically generated

H0: Null Hypothesis would be that there is no difference in outomes. To be more specific, I guess we could say that if we let pO be the probability of success for the Operative case, and pP be the probability of success for the Physical Therapy case, then the Null hypothesis expects these to be the same, i.e., pp = pO. This isn’t enough information to determine the probability of recovery. So we’d probably further presume to say that pp = pO = 110/200 = 0.55.

HA: Alternative Hypothesis is that there is a statistically significant difference in outcomes, i.e., pO ≠ pp. And of course we’d suspect that the Operative treatment is superior.

Test: So to test, we want to presume Null hypothesis is true, that pO = pP ≈ 0.55, and calculate the probability distribution, Pη0(m), which is the probability distribution for the difference in success ratios m = pO – pP for a random sample of 100 drawn from the Operative group and 100 drawn from the Physio group. But what is the probability distribution Pη0(m)? So let the measurement random variable be:



where X1 is the number of positive outcomes from the Operative group and X2 the number of positive outcomes from the Physio group. X1 and X2 are presumed binomially distributed,



and n1 = n2 = 100, and p = 0.55. But not sure what we can say about the probability distribution of the sum (or difference) of two binomially distributed random variables. Better to approximate as normal distributions, which can do since n1,2 = 100 >> 30. So we can say that:



The probability distribution of the difference of two normally distributed variables is also normally distributed, with the following mean and variance,



Or we could look at the probability distribution function of the difference of *proportions*. This is probably more to the point. So let the measurement random variable be:



and then according to our hypothesis that the proportions are the same so that <X>/n1 = <X2>/n2 = p, we have:



And again, we’d have to estimate p ≈ 110/200. So,



and therefore,



Now our measurement



And we want to know



Well we can see that m\* = 0.14 is 2 std from the mean. So,



So this is right at the threshold. So maybe still accept the Null hypothesis. We can do this with statsmodels. And note, we don’t enter in a Null Hypothesis value = v here, we just let it work out for itself that the Null Hypothesis value is the overall proportion.

Within a little rounding error, we see we get the same thing.



**Example: χ2 Test**

Let’s revisit the example above. Consider stats on two treatment options for proximal humerous fractures. We’d like to know if there is a statistically significant difference in outcomes. But this time, we’ll use the χ2 test.

A picture containing graphical user interface

Description automatically generated

H0: Null Hypothesis would be that there is no difference in outomes. So we presume the events are independent. Then the predicted numbers would be:



And,



And this would result in the following prediction,

|  |  |  |  |
| --- | --- | --- | --- |
|  | Improved | Not-Improved | rt |
| Operative | 55 | 45 | 100 |
| Physio | 55 | 45 | 100 |
| ct | 110 | 90 | 200 |

HA: Alternative Hypothesis is that there is a statistically significant difference in outcomes, i.e., pO ≠ pp. And of course we’d suspect that the Operative treatment is superior.

Test: So to test, we want to presume Null hypothesis is true, that pO = pP ≈ 0.55, and calculate the probability distribution, Pη0(m), which is the probability distribution of Z2 for a random sample of 100 drawn from the Operative group and 100 drawn from the Physio group. Gotta calculate the statistic,



And there is ν = 1 d.o.f. So our p-value would be:



which is what we got previously, doing it the other way.

**Example: Rank-Sum Test (Nonparametric)**

Now let’s look at some data we can apply non-parametric tests to in regards testing for a difference in means. I’m going to use the same data as for the matched pair experiment below, just to see the difference in outcome between the matched pair-mean tests and independent random samples mean tests. So let’s start with our data,

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

We would suspect that Group 2 > Group 1. Either way, we start by ranking the data from low to high,

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 4 | 7 |
| 6 | 2.5 |
| 2.5 | 8 |
| 1 | 5 |

Then we add up the ranks.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 13.5 | 22.5 |

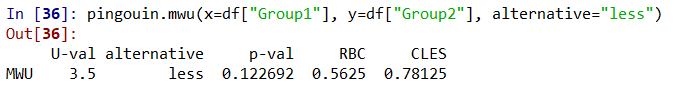
Now there are eight ways to make Group 2’s rank higher (or same). Here’s the possibilities, including the one we have. Note the second column is us swapping out our original 2.5 for the other 2.5.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group 2 |  |  |  |  |  |  |
| 7 | 2.5 | 2.5 | 4 | 4 | 4 | 5 |
| 2.5 | 5 | 6 | 5 | 5 | 6 | 6 |
| 8 | 7 | 7 | 6 | 7 | 7 | 7 |
| 5 | 8 | 8 | 8 | 8 | 8 | 8 |

So then the probability of achieving a higher rank sum, if the means are indeed equal is:



We’ll observe that this is lower than what we get for matched pair difference in mean tests. So that seems to make a real difference. The Mann-Whitney Test is supposed to be equivalent to this. And we can run it in pingouin.



But it clearly returns something a little different.

**Example: Kruskal-Wallis Test**

Let’s analyze the same data using the Kruskal-Wallis Test. This will test whether there is a statistically significant difference in the medians of the two groups.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

and rank it from low to high,

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 4 | 7 |
| 6 | 2.5 |
| 2.5 | 8 |
| 1 | 5 |

And we calculate SSEBG,



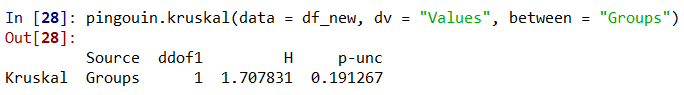
Then we form the statistic,



This statistic approximately follows a χ2 distribution with ν = G-1 = 2 – 1 = 1 d.o.f. So let’s see what p-value is associated with this.



Did this in pingouin.



So looks like our χ2 approximation wasn’t bad.

**Example: Matched Pair Tests**

Say we have matched pair group data.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

Let’s do a matched pair T-test for whether the means differ or not. So first we get the differences,

|  |
| --- |
| Group 2 – Group 1 |
| 5 |
| -5 |
| 9 |
| 3 |

And calculate the mean, and sample std. We get, according to numpy,



So our T-statistic is, under the Null Hypothesis that the populations are the same,



And is T-distributed with n-1 = 4 – 1 = 3 d.o.f. So to test for whether Group 2 is larger than Group 1, we’d calculate,



So at the customary 95% confidence level (p = 0.05), we would fail to reject the Null Hypothesis. Let’s revisit our table and do some nonparametric tests on it. We’ll start with the matched pairs sign test.

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

What is the p-value for the Alternative Hypothesis that group 2 has a larger mean according to the sign-test? Well we construct the table of signs,

|  |
| --- |
| Group 2 – Group 1 |
| 5 |
| -5 |
| 9 |
| 3 |

and we calculate the probability that we’d have as many positives as we do, or more. So,



So we’d provisionally still accept the Null Hypothesis that there is no difference. Turns out this is an abnormally high p-value compared to the other tests we’ll use below, and the Student’s T-test above. And I think this is because it doesn’t take into account the actual values of data differences, just the signs. Now let’s do the Wilcoxon signed rank test. So we take our diff

|  |  |
| --- | --- |
| Group 1 | Group 2 |
| 27 | 32 |
| 31 | 26 |
| 26 | 35 |
| 25 | 28 |

take the differences and rank them,

|  |
| --- |
| Group 2 – Group 1 Ranks |
| 5 (2.5) |
| -5 (2.5) |
| 9 (4) |
| 3 (1) |

Then we calculate the rank sum for positive and negative differences separately,

|  |  |
| --- | --- |
| Group 2 – Group 1 Ranks Sums | |
| T+ = 7.5 | T- = 2.5 |

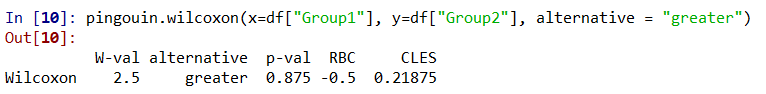
And as we said, T+ and T- follow a Wilcoxon pdf, which is approximately normal for large n. I doubt our n = 4 is considered large. But let’s just use it anyway. So letting x = T+ or T-, we have:



and we want to know the probability that Group 2 > Group 1. So we’ll look at T-. T- should be smaller than what the Null Hypothesis would predict if Group 2 > Group 1. So let’s calculate:



So the probability we’d get a T- as small as we do, given the Null Hypothesis, is 18%. So this indicates we don’t have evidence to reject the Null Hypothesis. Let’s do this in pingouin, and using the exact pdf appropriate for small n.



W is our T-. So we calculated that correctly. As expected, our p-value is off quite a bit, since n = 4 is too small to use the normal distribution approximation. It says p-val = 0.875, but I think this is to be interpeted as the confidence interval. So we can be 87.5% confidence the Alternative Hypothesis is correct, that Group 2 > Group 1. And the p-value defined, the likelihood of getting our W given the Null Hypothesis is correct is 1 – 0.875 = 0.125. This is more in line with our prior estimate of 0.18.

**Appendix**

We showed in the Descriptive Statistics file that:



is smallest for μ´ = μ = N-1Σixi. So this suggests a method for determining if two drugs, A and B, are different. We can do a histogram of recovery times for A and one for B. And calculate A, sA and B, sB. And then we can do a histogram of the total, and calculate the standard deviation of it, T, sT=A+B. If the two drugs are more or less identical, then sT should be < sA, sB. But if they are identical, then we should find sA, sB < sT rather.