**Hypothesis Tests for Independence**

Now I’d like to consider hypothesis tests regarding correlations between groups. This has to do with whether the two groups/columns of data vary in the same direction or not. Alas, the ‘varying’ it tests for is the linear sort. So if the data obey a perfect quadratic relationship, we will not get a perfect correlation coefficient. I think there are other correlation coefficients out there though, that redress this issue, and might be worth looking into.

**Null Hypothesis**

Our Null Hypothesis will often be, ‘there is no correlation between the groups’.

**Alternative Hypothesis**

The alternative hypothesis is that the groups are indeed correlated, of course.

**The tests**

There are various test statistics we can formulate. These depend on whether we’re comparing columns of numerical data, or of ranked categorical/numerical data, or of proportions.

**Correlations between numerical data**

If we have two columns of numerical data, X and Y,

|  |  |
| --- | --- |
| **X** | **Y** |
| x1 | y1 |
| x2 | y2 |
| x3 | y3 |
| x4 | y4 |
| x5 | y5 |
| x6 | y6 |
| x7 | y7 |
| x8 | y8 |

then we can define a correlation coefficient between them, the same as the one we introduced several files back:



Explicitly, this is:



And we’ll note these variances are using the population variance formula. Well, there is apparently a statistic that we can formulate to test hypotheses regarding the values of correlation coefficients. One can show that:



is a T-statistic, i.e., it follows a Student’s T distribution, where n is the number of data points, and ν = n-2 is the number of d.o.f. of the T distribution. So lettint t = x, we have:



**Correlations between ranked categorical data**

Say one or both of our columns consists of categorical data, which, however, can be ranked in some sort of order. An example would be, say, SAT scores, and Attractiveness. The SAT score column is numerical. Attractiveness is presumably a categorical ranking from lowest attractiveness to highest: 1, 2, 3, 4, …, n. So to compare SAT and Attractiveness, we’d similarly rank the SAT scores in order from lowest to highest: 1, 2, 3, 4. Note that if any of the unranked categorical values are identical, then we’d give them each the average of the two rankings they’d otherwise occupy. So then we’d have our two columns of ranked data. x1…8 would range from 1 to 8 in no necessary order, and likewise, y1…8 would run from 1 to 8 in no necessary order.

|  |  |
| --- | --- |
| **X** | **Y** |
| x1 | y1 |
| x2 | y2 |
| x3 | y3 |
| x4 | y4 |
| x5 | y5 |
| x6 | y6 |
| x7 | y7 |
| x8 | y8 |

And then we’d calculate the Spearman rank correlation coefficient rs, which is the same as the guy above.



Explicitly, this is:



I’m not sure if we can use the same test statistic as above. Perhaps?

**Correlations between proportions: χ2 Test**

We can use the ever-versatile χ2 test to examine the case for real differences in proportions in categorical data. For instance we could have students classified by SAT scores ranges, and pass/fail rates for a class. And we’d like to know whether the proportions in the classes are correlated.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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. In terms of probabilities, we want to know whether P(j|ℓ) = P(j)P(ℓ) or not, where j = 1,2, and ℓ = a, b, c. If they are equal, then the events are independent, and being in category 1, or 2, has no bearing on whether you are category a, b, or c. Otherwise, the events are correlated. So our Null Hypothesis is that the events are uncorrelated. Therefore we need to estimate P(j) and P(ℓ). Well, we’d say,



And therefore,



And we can get the predicted occupation numbers by multiplying the probabilities by the total number of entries, n. So



All that’s left is to determine ν. The number of d.o.f. of the χ2 distribution will be:



This is because presuming the j events are independent of the ℓ events, we have jmax – 1 d.o.f. to choose the values of p(j), and ℓmax – 1 d.o.f. to choose the values of p(ℓ). Because the probabilities of a given event are arbitrary except for the fact that they must add to 1.



follows a chi2 distribution (with x = Z2)



**Fisher’s Exact Test**

This is similar to the χ2 test, but is more accurate for smaller sample sizes than the χ2 test. So consider again that we run two experiments: 1, 2, each with two outcomes a, b (haven’t seen an example with more than two outcomes yet). And that we get the following results, njℓ.

|  |  |  |
| --- | --- | --- |
|  | a | b |
| 1 | n1a | n1b |
| 2 | n2a | n2b |

Then presuming events 1,2 are independent of the events a,b, the probability of getting the results we do is given by the hypergeometric distribution:



where n = n1a + n1b + n2a + n2b. Letting n1 = n1a + n1b, n2 = n2a + n2b, and na = n1a + n2a, nb = n1b + n2b, and n = n1 + n2 = na + nb, we can write this as:



The numerator is the number of ways to select n1a terms out of na, and n1b terms out of nb. The denominator is the number of ways to segregate the n terms into disjoint sets of n1 and n2 terms respectively.

**Example**

An example from the book is Teacher Certification Exam score and Teaching Presentation Score. They present the following table,

|  |  |  |
| --- | --- | --- |
| **Teacher** | **Presentation** | **Exam** |
| 1 | 7 | 44 |
| 2 | 4 | 72 |
| 3 | 2 | 69 |
| 4 | 6 | 70 |
| 5 | 1 | 93 |
| 6 | 3 | 82 |
| 7 | 8 | 67 |
| 8 | 5 | 80 |

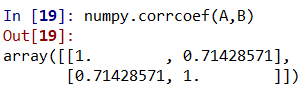
We’d like to know if there is a correlation or not between Rank and Score. So what we do is convert the numerical column to a categorical column, ranking all the scores from lowest (higher rank perhaps), to highest (lower rank, I guess). If two exam scores were same, and if they were to occupy ranks 6, and 7, say, then we’d give both scores the rank of 6.5. Anyway, converting our Exam column to a rank column, we have:

|  |  |  |
| --- | --- | --- |
| **Teacher** | **Presentation** | **Exam** |
| 1 | 7 | 8 |
| 2 | 4 | 4 |
| 3 | 2 | 6 |
| 4 | 6 | 5 |
| 5 | 1 | 1 |
| 6 | 3 | 2 |
| 7 | 8 | 7 |
| 8 | 5 | 3 |

and then we fill the numbers into our formula,



We can get this more directly from numpy using,



where A and B were defined as the respective columns.

**Example**

Let’s do an example from the book. They give us these numbers for number of shots and whether one subsequently got the flu or not. And we’d like to determine whether the number of flu cases is at all dependent on the number of shots.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 shots | 1 shot | 2 shots |
| Flu | 24 | 9 | 13 |
| Not-Flu | 289 | 100 | 565 |

It does seem, prima facie, that the number of shots makes a difference. The proportion of flu/not-flu for 0 shots, 1 shots, 2 shots is 24/289, 9/100, 13/565. So the proportion is decreasing. But let’s verify with the χ2 test. First we need to find the row and column totals,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 shots | 1 shot | 2 shots | row totals |
| Flu | 24 | 9 | 13 | 46 |
| Not-Flu | 289 | 100 | 565 | 954 |
| col totals | 313 | 109 | 578 | 1000 |

and then construct the predicted numbers.



So our predicted values are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 shots | 1 shot | 2 shots | rt |
| Flu | 14 | 5 | 27 | 46 |
| Not-Flu | 299 | 104 | 551 | 954 |
| ct | 313 | 109 | 578 | 1000 |

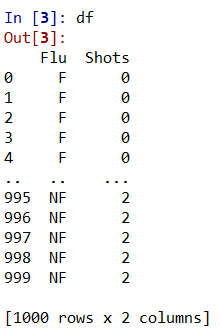
And our d.o.f. are ν = (2-1)(3-1) = 2. Next we form Z2,



And then we calculate the p-value of having an Z2 at least as large as this,



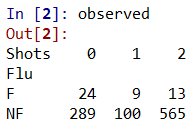
So we would reject the Null Hypothesis that they’re uncorrelated. We can do this in pingouin. I created a dataframe with our values. Looks like this:



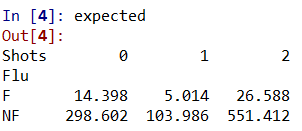
And the output the observed and expected contingency tables,



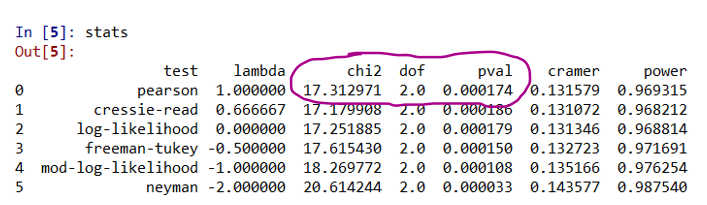
Here’s the observed table, which duplicates what we started with above,



and the expected table,



which also matches. And then the statistics. I think just the top row applies to us presently,



Can see the Z2\* value they get is 17.3, and the p-value is 0.000174. These values are a little different than ours because of rounding, which makes a larger difference in this case than you’d expect.

**Example**

The supreme court is 6-3 conservative, and on a recent court case, they voted 6-3 in favor. What are odds these events aren’t correlated?

Contigency table is:

|  |  |  |  |
| --- | --- | --- | --- |
|  | yes | no | rt |
| R | 6 | 0 | 6 |
| D | 0 | 3 | 3 |
| ct | 6 | 3 | 9 |

Expected values would be this. There is a 6/9 = 2/3 chance of being Republican, and a 6/9 = 2/3 chance of voting yes. If this are independent, then probability of the four outcomes is: P(RY) = 4/9, P(RN) = 2/9, P(DY) = 2/9, PR(DN) = 1/9. And we find the expectation values by multiplying by N = 9. Or, using our general formula from above, where we just multiply together the sum of the event’s respective row and column,



Which would look like,

|  |  |  |  |
| --- | --- | --- | --- |
|  | yes | no | rt |
| R | 4 | 2 | 6 |
| D | 2 | 1 | 3 |
| ct | 6 | 3 | 9 |

Z2 statistic is:



And there is ν = (2-1)(2-1) = 1 d.o.f. The p-value works out to:



So odds of getting that correlation are pretty small, if random. Let’s try a Fisher Exact test.



The Fisher test is supposed to be more accurate for small sample sizes, say, I think we’d go with this number, rather.