20.0 Assessing Significance Tests

- Answer Questions
- Review of Goodness-of-Fit
- Tests of Independence
- Simpson's Paradox

20.1 Review of Goodness-of-Fit

Recall that in this type of test, the null and alternative do not follow the usual pattern. They are:

$$H_0\text{: The model holds} \qquad \text{vs.} \qquad H_A\text{: The model fails.}$$

For a specific model, one gets more specific. For Darwin's peony experiment, the hypotheses were:

$$H_0\text{: The ratios of red, white and pink are } 1/4\text{: } 1/4/\text{: } 1/2$$
$$H_A\text{: The ratios differ from } 1/4\text{: } 1/4/\text{: } 1/2.$$

Note that we can only reject the model. We cannot prove it, since we never "prove" the null hypothesis. All we do is fail to reject it.

The test statistic for the goodness-of-fit test is:

$$ts = \sum \frac{(O_i - E_i)^2}{E_i}$$

where the sum is taken over all categories. The O_i is the observed count in category i, and E_i is the count predicted in that category by the model.

The significance probability comes from a chi-squared table. Let W be a chi-squared random variable with $k = \#\text{categories} - 1$ degrees of freedom.

For the Darwin example, the test statistic was 1.02 and we referred this to a chi-squared distribution with $k = 2$ degrees of freedom. The significance probability was:

$$P - value = \mathbf{P}[W \geq ts] = \mathbf{P}[W \geq 1.02].$$

From the table, this is between .7 and .5. So the null is not rejected.

20.2 Tests of Independence

Tests of independence are very similar to the goodness-of-fit tests. This is no accident, since independence is a specific kind of model.

Often one has a sample of cases; each case can be categorized according to two different criteria:

- each person got the drug or got a placebo, and each person lived or died;
- a criminal got the death penalty or not, and the state (AL, AZ, ...) in which they were charged;
- letter grade in a statistics course, and major

Suppose one took the 50 U.S. states and classified them as to whether they supported Bush or Kerry, and how many executions they had in the last five years (e.g., 0, 1-5, more than 5). You might get a contingency table that looks like this:

	Kerry	Bush	
0	20	1	21
1-5	5	8	13
> 5	2	14	16
	27	23	50

Here there are 20 states that supported Kerry and had no executions, 1 state that supported Bush and had no executions, etc.

The general null and alternative hypotheses are:

H_0: The two criteria are independent. H_A: Some dependence exists.

For a specific situation, it is better to be more specific. For this example, the hypotheses are:

H_0: Voting preference has nothing to do with execution rates.
H_A: There is a relationship between voting choice and executions.

Now we need to get a test statistic. The strategy is much like that for goodness-of-fit tests.

The test statistic is

$$ts = \sum_{\textbf{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

The O_{ij} is the observed count for the cell in row i, column j.

The E_{ij} uses the following formula:

$$E_{ij} = \frac{(i\textbf{th row sum}) * (j\textbf{th column sum})}{\textbf{total}}$$

This is why in the example contingency table I showed the row sums, the column sums, and the total count—to make the calculation of E_{ij} more easy.

For our example, we find:

$$\begin{aligned}
E_{11} &= 21 * 27/50 = 11.34 \\
E_{12} &= 21 * 23/50 = 9.66 \\
E_{21} &= 27 * 13/50 = 7.02 \\
E_{22} &= 23 * 13/50 = 3.38 \\
E_{31} &= 27 * 16/50 = 8.64 \\
E_{32} &= 23 * 16/50 = 7.36
\end{aligned}$$

Then the test statistic is:

$$\begin{aligned}
ts &= \frac{(20-11.34)^2}{11.34} + \frac{(1-9.66)^2}{9.66} + \ldots + \frac{(14-7.36)^2}{7.36} \\
&= 32.29.
\end{aligned}$$

We compare the test statistic to the value from a chi-squared distribution with degrees of freedom equal to

$$k = \textbf{(number of rows - 1)} * \textbf{(number of columns - 1)}.$$

For our example, $k = (3 - 1) * (2 - 1) = 2$.

The significance probability is

$$P\textbf{-value} = \mathbf{P}[W > ts]$$

where W has the chi-squared distribution with k degrees of freedom.

For a chi-squared random variable with 2 degrees of freedom, the 1% value is 9.21. So

$$.01 = \mathbf{P}[W > 9.21] > \mathbf{P}[W > 32.29] = P\textbf{-value}.$$

So the significance probability is much less than .01. There is strong evidence that political preference and execution rates are somehow connected.

But the connection can be very subtle. We cannot infer causation, and the apparent relationship may not be at all what we expect.

Sometimes there are hidden confounders that are more interesting than the relationship between the two classification criteria. It can even happen that the hidden confounder can reverse the apparent relationship in the data. When this happens, it is called Simpson's Paradox.

20.3 Simpson's Paradox

Contingency tables can more more than two classification criteria (i.e., the row and column criteria). Sometimes the cases are classified according to three or more criteria; e.g., gender, major, and year could be used to classify a sample of students.

Sometimes the effect of third criterion changes the story. We saw this in the Berkeley admissions data. Using just the criteria of admission status and gender suggested sex-bias against women. When a third criterion, major, was added, this bias disappeared. The major was a confounding effect.

Between 1967 and 1977 there was a moratorium on the death penalty in the U.S. A main reason for this was the belief that there was racial bias in the application of the death penalty.

When the racial bias argument was tested in court, defenders of the death penalty produced a contingency table:

	Black	White	
death sentence	17	19	36
no death sentence	149	141	290
	166	160	326

According to this table, it looks slightly better to be Black than White when charged with murder. The chance of a Black person being sentenced to death is $17/(17+149) = .1024$, whereas the chance of a White person being executed is $19/(19+141)=.11875$.

Therefore death penalty supporters argued against the moratorium, claiming that there was no racial discrimination in the use of the death penalty.

How might a liberal do a hypothesis test to study the argument by the death penalty supporters?

$$\mathrm{H}_0\text{: } p_B - p_W \le 0 \text{ vs. } \mathrm{H}_A\text{: } p_B - p_W > 0.$$

The test statistic is

$$\begin{aligned} ts &= \frac{\hat{p}_B - \hat{p}_W}{\sqrt{\frac{\hat{p}_B(1-\hat{p}_B)}{n_B} + \frac{\hat{p}_W(1-\hat{p}_W)}{n_W}}} \\ &= -.4705. \end{aligned}$$

The significance probability is

$$P - \text{value} = \mathrm{P}[z \ge -.4705] = .674$$

So there is no evidence against the null hypothesis. It looks like the liberals lose.

But what about the effect of a confounding variable? Suppose we consider the race of the victim.

If the victim is white, then

	Black	White	
death sentence	11	19	20
no death sentence	52	132	184
	63	151	204

From this, the chance of death for a Black person is 11/(11+52)=.1746, while the chance of death for a White person is only 19/(19+132)=.1258. It hurts to be Black when the victim is White.

Similarly, if the victim is Black, then the table is

	Black	White	
death sentence	6	0	6
no death sentence	97	9	106
	103	9	112

A Black person has a chance of about $6/(6+97)=.05824$ of being executed, while a White person seems to have essentially a zero chance.

This is Simpson's Paradox. Race of the victim is the confounding variable. When that is considered, the liberals are correct—Blacks are more likely to be executed than Whites.