Stat 200: Introduction to Statistical Inference **Winter 2023**

Lecture 14: Categorical data

Lecturer: Dominik Rothenhäusler *March 6*

Note: *These lecture notes were written by Art Owen. If you like the material, he gets the credit! These notes have not been subjected to the usual scrutiny reserved for formal publications. They are meant as a memory aid for students who took stat 200 at Stanford University. They may be distributed outside this class only with the permission of Art Owen. Also, Stanford University holds the copyright.*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

14.1 Introduction

This lecture was about cross classified data based on Chapter 13 of Rice. It is easiest to visualize in terms of either people in categories or balls in urns or bins. The data are a list of integer counts $n_{ij} \in \{0, 1, 2, \dots\}$ for $i = 1, \dots, I$ and $j = 1, \dots, J$. In class I spoke of n_{ij} people of occupation i from country j. We need to have $I \geqslant 2$ and $J \geqslant 2$ for this to really be a cross-classification.

These notes explain the goals and reasoning. Be sure to read the numerical examples in the book.

The total number of people with occupation i is $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$, the total number of people from country j is $n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$ and the total number of people counted in the data is

$$n_{\bullet\bullet} = \sum_{i=1}^{I} n_{i\bullet} = \sum_{j=1}^{J} n_{\bullet j} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}.$$

The observed data look like

n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\bullet}$
n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n_{I1}	n_{I2}	$\cdots$	n_{1J}	$n_{I\bullet}$
$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	n_{1J}	$n_{\bullet\bullet}$

You can get a very small table of just $I \times J$ numbers even though the number $n_{\bullet\bullet}$ of observations behind it can be enormous. (This should remind you of sufficient statistics.) The null hypothesis for this data is that the row a person belongs to has nothing to do with the column that they belong to. Or if balls are placed within an $I \times J$ grid of bins, the row that a ball occupies has nothing to do with the column that it occupies.

Where the methods differ is in what is random. Each n_{ij} is the observed value of some random variable N_{ij}. Sometimes the row totals are fixed and sometimes they are random. Same for the column totals. Once we decide what is random and what is not, we can formulate the null hypothesis mathematically.

14.2 Fisher's exact test

Suppose that $I = J = 2$ and that $n_{i\bullet}$ is non random for $i = 1, 2$ and $n_{\bullet j}$ is non random for $j = 1, 2$. The classic example of this kind is about tea tasting.[1] Muriel Briston was a colleague of Ronald Fisher. She claimed that she could tell by taste the difference between a cup of tea where the tea had been added to the milk and a cup where the milk had been added to the tea. Maybe there is something noncommutative there where tea + milk $\neq$ milk + tea. Apparently Fisher was skeptical. If you think of what happens to the temperature of the milk over time it is plausible. If you drop a bit of milk into near boiling tea the milk is heated much more suddenly than if you pour tea into milk. Then again maybe the effect is not strong enough to be detectable. On the other hand, maybe some people are sensitive enough to detect a thing that other people cannot. This uncertainty calls for statistical testing.

Here's how you could test it. Make up 8 cups of tea, four done each way, with the expert not present. Then ask the expert to say which four got milk first. You get a table like

	Milk first chosen	Tea first chosen	Total
Milk was first	x	$4-x$	4
Tea was first	$4-x$	x	4
Total	4	4	8

which we could write as

N_{11}	N_{12}	$n_{1\bullet}$
N_{21}	N_{22}	$n_{2\bullet}$
$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

The row totals are fixed by the experimental design and hence not random. Same for the column totals. The table entries are random given the row and column totals.

A null hypothesis is that the taster simply guesses $n_{\bullet 1}$ cups at random in a way that has nothing to do with what is in the cup. Under that model there are $\binom{n_{\bullet\bullet}}{n_{\bullet 1}}$ equally probable choices. Here $\binom{8}{4} = 70$ different ways that Muriel Briston could have selected 4 cups that had milk before tea. Then N_{11} has the hypergeometric distribution,[2]

$$\Pr(N_{11} = n_{11}) = \frac{\binom{n_{1\bullet}}{n_{11}}\binom{n_{2\bullet}}{n_{21}}}{\binom{n_{\bullet\bullet}}{n_{\bullet 1}}} = \frac{\binom{n_{1\bullet}}{n_{11}}\binom{n_{2\bullet}}{n_{\bullet 1}-n_{11}}}{\binom{n_{\bullet\bullet}}{n_{\bullet 1}}}.$$

See the Wikipedia page for the legal outcome values (i.e., the support) of a hypergeometric distribution.

The hypergeometric distribution is available to you in R (`rhyper, dhyper, phyper, qhyper`). We can easily come up with p-values for the alternative that N_{11} is larger than predicted. E.g., better than random taste testing. It is

$$p = \Pr(N_{11} \geqslant n_{11}) = \sum_{x : x \geqslant n_{11}} \frac{\binom{n_{1\bullet}}{x}\binom{n_{2\bullet}}{n_{\bullet 1}-x}}{\binom{n_{\bullet\bullet}}{n_{\bullet 1}}}.$$

Getting the milk+tea question right for all 8 cups would give a one tailed $p = 1/70$. Similarly, $p = \Pr(N_{11} \leqslant n_{11})$ is a p-value for H_0 against an alternative that N_{11} is smaller than expected. For a two-sided alternative we use the hypergeometric distribution to get $\Pr\big(|N_{11} - \mathbb{E}(N_{11})| \geqslant |n_{11} - \mathbb{E}(N_{11})|\big)$. We need $\mathbb{E}(N_{11})$ to do that. It is $n_{1\bullet}n_{\bullet 1}/n_{\bullet\bullet}$. These are the p-values from Fisher's exact test.

Sometimes people use Fisher's exact test even when the row and column totals were not really fixed. Rice gave an example of a company with 24 male and 24 female employees of which 35 were promoted and 13

[1] `https://en.wikipedia.org/wiki/Muriel_Bristol`

[2] Rice p42 and `https://en.wikipedia.org/wiki/Hypergeometric_distribution`

not. The hypergeometric argument would apply directly if that company had decided to promote 35 of their employees.

This next bit is for your curiosity and not something you'll be tested on. Statisticians are still arguing about it. Sometimes Fisher's test is used even when the row and column totals were random. The argument is like the one in regression. If you see all the x_i you still know nothing about the slope β_1 until you get Y_i given x_i. For regression we made that precise by exhibiting a likelihood where the MLE and likelihood ratio for β_1 all came from the Y given x distribution.

For 2×2 tables, the intuitive argument is as follows. Suppose you see

?	?	$n_{1\bullet}$
?	?	$n_{2\bullet}$
$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet\bullet}$

with ? indicating an unknown value. Do you think you know anything about whether rows are associated with columns from that table? If not, then you might reason that the association comes only from the distribution of what is inside the table given the row and column sums. Intuition like this is good for formulating questions, for guessing answers, and for understanding an answer that the math gives, but intuition itself is not the answer. Rice avoids this issue and so will we!

Another off the record fact: There are generalizations of Fisher's exact test to $I \times J$ tables for any $I \geqslant 2$ and $J \geqslant 2$. The null hypothesis is that all possible tables with the given row and column sums have the same probability. The distribution of counts is then more complicated than hypergeometric but can be computed.

14.3 χ^2 test of homogeneity

Suppose that we fix the column totals. We observe $n_{\bullet j}$ people for $j = 1, \ldots, J$. The data in each of our J columns are independent. From column j we get multinomial counts $N_{1j}, \ldots, N_{Ij}$. The sample size in that column is $n_{\bullet j}$ and the row probabilities are $\pi_{i|j} \geqslant 0$ with $\sum_i \pi_{i|j} = 1$. [Rice uses π_{ij} here but I like to show that these are conditional probabilities.]

We can model row variable i having nothing to do with the column variable j through

$$H_0 : \pi_{i|1} = \pi_{i|2} = \cdots = \pi_{i|J}, \quad i = 1, \ldots, I.$$

Imagine that your chance of having occupation i in country j is the same for all j. The alternative is that $\pi_{i|j}$ can be different for every j.

We will do a generalized likelihood ratio test. Let the common value of $\pi_{i|j}$ under H_0 be called π_i. The null has $I-1$ degrees of freedom because these add up to 1. The alternative has $I-1$ degrees of freedom in each of J columns so it has $(I-1)J$ degrees of freedom. Therefore the GLRT will have

$$(I-1)J - (I-1) = (I-1)(J-1)$$

degrees of freedom.

To do the GLRT we need the MLEs under H_0 and H_A. Under H_0 $\hat{\pi}_i = n_{i\bullet}/n_{\bullet\bullet}$ as we saw in class and as Rice proves. Under H_A, $\hat{\pi}_{i|j} = n_{ij}/n_{\bullet j}$ because we have J separate multinomial likelihoods.

After the algebra is over, the likelihood ratio test statistic ends up as

$$-2\log(\Lambda) = 2\sum_{j=1}^{J}\sum_{i=1}^{I} O_{ij}\log(O_{ij}/E_{ij}).$$

The same Taylor approximation we used for multinomials before gives

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

If $X^2 \geqslant \chi^{2,1-\alpha}_{(I-1)(J-1)}$ then we reject H_0 at level α.

Here O_{ij} is the observed count N_{ij} in row i and column j while E_{ij} is the expected count there under H_0. Writing it this way makes it look just like our previous GLRT for the multinomial except now we use a double sum. We need to work out E_{ij}. Here

$$E_{ij} = n_{\bullet j} \times \hat{\pi}_{i|j} = n_{\bullet j} \times \frac{n_{i\bullet}}{n_{\bullet\bullet}} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}}.$$

Suppose we had gotten the exact same N_{ij} but we had instead sampled $n_{i\bullet}$ observations in row i for $i = 1, \ldots, I$ independently. E.g., we sample $n_{i\bullet}$ people with occupation i and record which country j they are in. We would end up with the exact same $-2\log(\Lambda)$ and X^2 as above and the exact same degrees of freedom.

14.4 χ^2 test of independence

Now suppose that $n_{\bullet\bullet}$ is fixed. We sample that many people and then record both their occupation and country getting a table of N_{ij} values. Or we drop balls from the sky and they land in an $I \times J$ grid of bins.

We now have one big multinomial distribution with IJ levels and probability π_{ij} on the (i, j) bin. A natural way to form a null hypothesis is to suppose that

$$\pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j}$$

where $\pi_{i\bullet} = \sum_{j=1}^{J} \pi_{ij}$. Then the row you land in is independent of the column.

The null hypothesis has $I - 1$ degrees of freedom for $\pi_{i\bullet}$ and $J - 1$ more for $\pi_{\bullet j}$ for a total of $I + J - 2$. The alternative hypothesis has IJ parameters π_{ij} that add up to one so it has $IJ - 1$ degrees of freedom. Therefore the GLRT has

$$(IJ - 1) - [(I - 1) + (J - 1)] = (I - 1)(J - 1)$$

degrees of freedom, the same as for testing homogeneity.

You can write out the likelihood under H_0. It ends up as the product of two multinomial likelihoods, one for rows and one for columns and we get

$$\hat{\pi}_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}}, \quad \text{and} \quad \hat{\pi}_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}.$$

Under H_A we have one combined multinomial on IJ levels. The MLE for that multinomial is just

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}.$$

Once again the likelihood ratio test ends up as

$$-2\log(\Lambda) = 2\sum_{j=1}^{J}\sum_{i=1}^{I} O_{ij} \log(O_{ij}/E_{ij}).$$

and Pearson's χ^2 is

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

both with with $O_{ij} = N_{ij}$ except that now $E_{ij} = \mathbb{E}(N_{ij}|H_0)$ is expectation under row and column independence. That turns out to be

$$E_{ij} = n_{\bullet\bullet}\times\hat{\pi}_{i\bullet}\times\hat{\pi}_{\bullet j} = n_{\bullet\bullet}\times\frac{n_{i\bullet}}{n_{\bullet\bullet}}\times\frac{n_{\bullet j}}{n_{\bullet\bullet}} = \frac{n_{i\bullet}n_{\bullet j}}{n_{\bullet\bullet}}.$$

This is the same as we got under the test for homogeneity.

There is really just that one test for all three scenarios.

14.5 Simpson's paradox

I presented some extreme hypothetical outcomes for two doctors. For Dr. B., the patient outcomes are

Dr. B	Lived	Died	Total
Sick	25	25	50
Well	50	0	50
Total	75	25	100

while Dr. K has

Dr. K	Lived	Died	Total
Sick	0	10	10
Well	80	10	90
Total	80	20	100

We see evidence of better outcomes for Dr. B on both patient groups. Pooling the data we get

	Lived	Died	Total
Dr. B	75	25	100
Dr. K	80	20	100
Total	155	45	200

Pooling the data makes it appear that Dr K is better. What went wrong is that Dr B was getting the more difficult cases and the pooled data do not show that.

The pooled data collapses a $2\times2\times2$ table down to a 2×2 table. Collapsing a table can make the opposite effect appear. Or it can create an association out of individual tables that don't have any. Or hide an association that is present in the individual tables.

People comparing doctors or hospitals take care to adjust for the varying case loads. A hospital with a strong reputation might attract the most difficult cases and then not look as good as it should in a collapsed table. If collapsed tables were used it could give surgeons an incentive to avoid the most difficult cases. A ride hailing driver might avoid night time if that is when harder to please riders arrive.

In class I mentioned a genetic marker that is associated with diabetes in the population at large but not in any of the subgroups. It is thus unlikely to be the key to any cure or treatment. The marker could be genuinely very predictive of diabetes just because it is predictive of being in the higher risk subgroup.

Given data it is easy to collapse it over one or more additional variables. The much more difficult problem is uncollapsing a table. You would have to think of which other variable might be important. That would

come from real world understanding of the nature of the rows and columns. Then you might find that the variable you're thinking of is not in anybody's data set.

14.6 Prospective vs retrospective

Suppose we look at the association of 2 diets and 2 health outcomes (good or bad). We could find 100 people with each diet and follow them prospectively for 50 years to measure the health outcome. Apart from the delay there is also the issue that if the bad outcome is rare we might get 0 out of 100 in each group (or some other small number) and have an uninformative comparison.

In a retrospective study, we find 100 people with the good outcome and 100 with the bad outcome and then try to ascertain what their past diet was. We are then assured of sufficient numbers in the two outcome groups. It is remotely possible that one of the dietary habits was extremely rare but we would probably know that before doing the study. There is also a practical difficulty in getting good data about the past (memory being imperfect).