

Ling 235 – Statistics, Jan 26, 2005

Today we'll be covering a number of topics:

- Contingency tables: introduction
- Odds ratios
- Counting, a bit of combinatorics, Binomial distribution
- Intro to hypothesis testing, degrees of freedom
- G^2 test (likelihood ratio)
- Chi-squared test
- Fisher's exact test

1. Contingency tables

There are many situations in quantitative linguistic analysis where you will be interested in the possibility of association between two categorical variables. In this case, you will often want to represent your data as a contingency table. Here's an example from my own research:

Parallelism in phrase coordination, [[NP1] and [NP2]]. I was interested in whether NP1 and NP2 tended to be similar to each other. As one instance of this, I looked at the patterns of PP modification in the Brown and Switchboard corpora, and came up with contingency tables like this:

		Brown		Switchboard			
		hasPP	noPP	hasPP	noPP		
NP1	hasPP	95	52	NP1	hasPP	78	76
	noPP	174	946		noPP	325	1230

2. Odds and odds ratios

Given a contingency table of the form

$$\begin{array}{ccccc} & & Y & & \\ & & y_1 & y_2 & \\ X & x_1 & n_{11} & n_{12} & \\ & x_2 & n_{21} & n_{22} & \end{array}$$

one of the things that's useful to talk about is how the value of one variable affects the distribution of the other. For example, the overall distribution of Y is

$$freq(y_1) = \frac{n_{11}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \quad freq(y_2) = \frac{n_{12}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}}$$

Alternatively we can speak of the overall *odds* of y_1 versus y_2 :

$$\frac{freq(y_1)}{freq(y_2)} = \frac{\frac{n_{11}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}}}{\frac{n_{12}+n_{22}}{n_{11}+n_{12}+n_{21}+n_{22}}} = \frac{n_{11}+n_{21}}{n_{12}+n_{22}}$$

If $X = x_1$, then the odds for Y are just $\Omega_1^Y = \frac{n_{11}}{n_{12}}$. If the odds of Y for $X = x_2$ are greater than the odds of Y for $X = x_1$, then the outcome of $X = x_2$ **increases** the chances of $Y = y_1$. We can express the effect of the outcome of X on the odds of Y by the **odds ratio** (which turns out to be symmetric between X, Y):

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

An odds ratio $\theta = 1$ indicates no association between the variables. For the Brown and Switchboard parallelism examples:

$$\theta_{Brown} = \frac{95 \times 946}{52 \times 174} = 9.93 \quad \theta_{Swbd} = \frac{78 \times 1230}{325 \times 76} = 3.88$$

So the presence of PPs in left and right conjunct NPs seems more strongly interconnected for the Brown (written) corpus than for the Switchboard (spoken).

2. Counting, combinatorics and binomial distribution

Take a binary variable X , let its possible values be 0 and 1. (e.g., coin flip: heads=1, tails=0; NP conjunct: hasPP=1, noPP=0, ...) Call p its probability of outcome 1. (This is called a **Bernoulli random variable** with parameter p .)

Suppose we take a variable Y to be the sum of four such Bernoulli random variables X_i , where all the X_i are independent and each has the same parameter p . By counting, the probability distribution of Y is:

Y	$P(Y)$
4	$[\binom{4}{4}]1 \times p^4$
3	$[\binom{4}{3}]4 \times p^3(1-p)$
2	$[\binom{4}{2}]6 \times p^2(1-p)^2$
1	$[\binom{4}{1}]4 \times p(1-p)^3$
0	$[\binom{4}{0}]1 \times (1-p)^4$

This is called the four-trial ***binomial*** distribution with parameter p . A binomial distribution is determined by two parameters: the Bernoulli probability p , and the number of trials n . The binomial probability distribution for p, n is

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of ways of choosing k out of n items.

The generalization of a binomial distribution to random variables with more than two outcomes is called a ***multinomial*** distribution.

3. Degrees of freedom

This cool-sounding concept just means the number of free parameters that you estimate when fitting a model to your data. For example, if you wanted to fit a binomial distribution to a dataset, and you already knew the number of trials n , then there would be one free parameter p to fit, and hence one degree of freedom in your model.

As a more complicated example, suppose we wanted a model the coordination example above where PP modification was independent in left and right NP, and had the same likelihood in the two conjuncts. This would only have one parameter, hence one degree of freedom. In a slightly more general model, PP modification in the two conjuncts is independent, but the likelihood may differ by conjunct. This model would have two degrees of freedom. Note that the latter model space contains the former.

As another example, a multinomial distribution with k possible outcomes has $k - 1$ free parameters (since there is a constraint that all the p_i parameters sum to 1).

4. Introductory hypothesis testing.

Hypothesis testing typically goes like this: think up a specific, relatively simple model for your data. Call this the ***null hypothesis*** H_0 . Contrast that with a more general model for your data, of which H_0 is a special case. Call this more general model H_A , the ***alternative hypothesis***. You will then calculate some ***statistic*** of your data based on H_0 with respect to H_A . That statistic will have some probability distribution on the assumption that H_0 is correct. If the value of the statistic likelihood of your data under H_0 and H_A . If the likelihood under H_0 is low enough, reject H_0 in favor of the more general H_A .

5. Likelihood ratio test.

With this test, the statistic you calculate for your data D is the ***likelihood ratio***

$$\Lambda^* = \frac{\max P(D; H_0)}{\max P(D; H_A)}$$

that is: the ratio of the maximum data likelihood under H_0 to the maximum data likelihood under H_A . This requires that you explicitly formulate H_0 and H_A . $-2 \log \Lambda^*$ is distributed like a chi-squared [see below] with ***degrees of freedom*** equal to the difference in the the number of free parameters in H_A and H_0 . [Danger: don't apply this test when expected cell counts are low, like < 5 .]

6. Chi-squared test.

Apply this test to all sorts of contingency tables, if you have a model with k parameters that predicts expected values E_{ij} for all cells. You calculate the X^2 statistic:

$$X^2 = \sum_{ij} \frac{[n_{ij} - E_{ij}]^2}{E_{ij}}$$

In the chi-squared test, H_A is the model that each cell in your table has its own parameter p_i in one big multinomial distribution. Therefore, if your contingency table has n cells, then the difference in the number of free parameters between H_A and H_0 is $n - k - 1$. Correspondingly, you can look up the p -value of your X^2 statistic for $n - k - 1$ ***degrees of freedom***.

[Danger: don't apply this test when expected cell counts are low, like < 5 .]

7. Fisher's exact test.

This test applies to a 2-by-2 contingency table:

		Y		
		y_1	y_2	
X	x_1	n_{11}	n_{12}	n_{1*}
	x_2	n_{21}	n_{22}	n_{2*}
	n_{*1}	n_{*2}		n

H_0 is the model that all ***marginal totals*** are fixed, but that the individual cell totals are not – alternatively stated, that the individual outcomes of X and Y are independent. [H_0 has one free parameter – why?] H_A is the model that the individual outcomes of X and Y are not independent. With Fisher's exact test, you directly calculate the *exact* likelihood of obtaining a result as extreme or more extreme than the result that you got. [Since it is an *exact* test, you can use Fisher's exact test regardless of expected and actual cell counts.]