Categorical data and contingency tables

Kerby Shedden

Department of Statistics, University of Michigan

October 22, 2014

Proportions

Suppose we have an independent and identically distributed (iid) sample $X_1, \ldots, X_n$ of binary responses. For example, each X_i may be an individiual's response to a yes/no question in a survey.

The distribution of each X_i is characterized by the "success probability"

$$p \equiv P(X_i = 1).$$

The mean and variance of each X_i are

$$EX_i = p \qquad \text{var}(X_i) = p(1-p).$$

The mean and variance of $\bar{X}$ are

$$E\bar{X} = p \qquad \text{var}(\bar{X}) = p(1-p)/n.$$

Confidence intervals for a proportion

We can form an approximate 95% confidence interval for p in the same way we would form a CI for the expected value EX:

$$\bar{X} \pm 2\hat{\sigma}/\sqrt{n} \qquad \text{or} \qquad \hat{p} \pm 2\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}.$$

Contingency tables

Suppose we have paired binary data X_i, Y_i. For example, X_i might indicate whether an individual is employed, and Y_i might indicate whether that same individual has completed at least two years of college.

Note that it is critical that this is cross-classified data – for each subject, we have data on the two variables of interest (as opposed to having separate samples of data for X and Y).

There are only four possible values for each pair,

$$(0,0), (0,1), (1,0), (1,1),$$

where $(0,0)$ indicates that the individual is not emplyed and has not completed two years of college, $(0,1)$ indicates that the individual is not employed and has completed two years of college, and so on.

Contingency tables

The information in our sample of n pairs can be summarized in a 2×2 contingency table

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	n_{11}	n_{10}
$X_i = 0$	n_{01}	n_{00}

Here, n_{11} is the number of subjects who have $X_i = Y_i = 1$, and so on.

Contingency tables

Here is an example of a 2×2 contingency table:

	$\geq$ 2 yr college	$<$ 2 yr college
Employed	74	65
Not employed	23	31

Log odds ratios

Suppose we are interesting in assessing whether there is any association between X_i and Y_i. For example, we may want to know whether it is more likely for college educated people to be employed (compared to what we would expect if college education and emplyment status were independent).

To do this, we could use the correlation coefficient, but for binary data it is more common to use the log odds ratio.

Log odds ratios

First, suppose we have a random variable Z with success probability p, i.e. $P(Z = 1) = p$, $P(Z = 0) = 1 - p$. The odds for Z are

$$P(Z = 1)/P(Z = 0) = p/(1 - p).$$

For example,

p	Odds
$1/2$	1
$1/3$	$1/2$
$2/3$	2

Log odds ratios

Based on this contingency table,

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	n_{11}	n_{10}
$X_i = 0$	n_{01}	n_{00}

we can estimate the odds for Y, restricted to the subpopulation with $X = 1$, as

$$\frac{P(Y = 1|X = 1)}{P(Y = 0|X = 1)} = \frac{n_{11}/(n_{11} + n_{10})}{n_{10}/(n_{11} + n_{10})} = n_{11}/n_{10}.$$

Log odds ratios

Similarly, the estimated odds for Y restricted to the subpopulation with $X = 0$ is

$$\frac{n_{01}/(n_{01}+n_{00})}{n_{00}/(n_{01}+n_{00})} = n_{01}/n_{00}.$$

The ratio between these odds is the sample odds ratio

$$\frac{n_{11}/n_{10}}{n_{01}/n_{00}} = \frac{n_{11}n_{00}}{n_{10}n_{01}}.$$

Note that the numerator of the odds ratio is the product of frequencies for the concordant cells (n_{11} and n_{00}), and the denominator is the product of frequencies for the discordant cells.

Log odds ratios

The sample log odds ratio is the natural logarithm of the sample odds ratio

$$\mathrm{LOR} = \log(n_{11}) + \log(n_{00}) - \log(n_{01}) - \log(n_{10}).$$

Log odds ratios

The sample log odds ratio is positive when there is more concordant data ($X = Y = 1$ or $X = Y = 0$). It is negative when there is more discordant data ($X = 1,\ Y = 0$ or $X = 0,\ Y = 1$).

Put another way, the sample log odds ratio is positive when n_{11} and n_{00} are big compared to n_{10} and n_{01}.

The sample log odds ratio is zero if and only if the conditional odds for Y when $X = 1$ is the same as the conditional odds for Y when $X = 0$. This is equivalent to X and Y being independent.

Log odds ratios

An important fact is that if we reverse the rolls of X and Y in calculating the odds ratio (i.e. if we divide the conditional odds for X given $Y = 1$ by the conditional odds for X given $Y = 0$), we get the same result.

Another important fact is that there is a simple approximate formula for the variance of the log odds ratio:

$$\mathrm{var}(\mathrm{LOR}) = 1/n_{11} + 1/n_{00} + 1/n_{01} + 1/n_{10}.$$

This makes it possible to construct Z-scores, hypothesis tests, and confidence intervals for the population log odds ratio centered at the sample log odds ratio.

Population odds ratios

The population odds ratio and log odds ratio for the joint probability table are

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	p_{11}	p_{10}
$X_i = 0$	p_{01}	p_{00}

$$\mathrm{OR} = \frac{p_{11}p_{00}}{p_{10}p_{01}}$$

$$\mathrm{LOR} = \log(p_{11}) + \log(p_{00}) - \log(p_{10}) - \log(p_{01})$$

An alternative expression for the variance of the sample log odds ratio is

$$n^{-1}\left(1/p_{11} + 1/p_{00} + 1/p_{10} + 1/p_{01}\right)$$

Odds ratio example

Suppose the following is a joint distribution, and the expected cell counts for a sample of size 50:

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	0.4	0.25
$X_i = 0$	0.25	0.1

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	20	12.5
$X_i = 0$	12.5	5

The population odds ratio is

$$\mathrm{OR} = \frac{0.4 \cdot 0.1}{0.25 \cdot 0.25} = 0.64$$

The population log odds ratio is

$$\mathrm{LOR} = \log(0.4) + \log(0.1) - \log(0.25) - \log(0.25) = -0.45$$

Odds ratio example

Now suppose we observe the following data set:

	$Y_i = 1$	$Y_i = 0$
$X_i = 1$	23	11
$X_i = 0$	10	6

The sample odds ratio and log odds ratio are

$$\widehat{\mathrm{OR}} = \frac{23 \cdot 6}{11 \cdot 10} = 1.25$$

$$\widehat{\mathrm{LOR}} = \log(23) + \log(6) - \log(11) - \log(10) = 0.22$$

The standard error of the log odds ratio is

$$\mathrm{SE}(\widehat{\mathrm{LOR}}) = \sqrt{1/23 + 1/6 + 1/11 + 1/10} = 0.63$$

Odds ratio example

A 95% confidence interval for the true log odds ratio is

$$\widehat{\mathrm{LOR}} \pm 2\mathrm{SE} = 0.22 \pm 1.26.$$

A p-value for the level 0.05 two-sided test of the null hypothesis that $\rho = 0$ is

$$2P(Z < -|\widehat{\mathrm{LOR}}/\mathrm{SE}|) = 2P(Z < -0.35) = 0.73.$$

Log odds ratios for several tables

Suppose we have K 2×2 tables, all of which are based on the same variables. The K tables might contain data of a similar type collected in K different populations.

Let n_{ijk} denote the count for the i, j cell in the k^{th} table, for $i, j = 1, 2$. The log odds ratio for this series of tables is

$$\hat{\theta}_{\text{MH}} \equiv \log \frac{\sum_k n_{11k} n_{22k} / N_k}{\sum_k n_{12k} n_{21k} / N_k}$$

This is called the "Mantel-Haenszel" estimator.

Note that $\hat{\theta}_{\text{MH}}$ is not the same as calculating the log odds statistic separately for each table and then averaging the statistics.

Log odds ratios for several tables

A famous data set from Agresti's book:

City	Smoking	Lung cancer Y	Lung cancer N	Smoking proportion	Cancer proportion	LOR
Beijing	Y	126	100	0.70	0.50	0.79
	N	35	61			
Shanghai	Y	908	688	0.55	0.48	0.76
	N	497	807			
Shenyang	Y	913	747	0.64	0.48	0.78
	N	336	598			
Nanjing	Y	235	172	0.69	0.50	1.04
	N	58	121			
Harbin	Y	402	308	0.68	0.50	0.84
	N	121	215			
Zhengzhou	Y	182	156	0.67	0.50	0.46
	N	72	98			
Taiyuan	Y	60	99	0.75	0.33	0.86
	N	11	43			
Nanchang	Y	104	89	0.77	0.50	0.69
	N	21	36			

Log odds ratios for several tables

The overall estimate of the association between smoking and lung cancer is:

$$\hat{\theta}_{\mathrm{MH}} = \log \frac{126 \cdot 61/322 + \cdots + 104 \cdot 36/250}{35 \cdot 100/322 + \cdots + 21 \cdot 89/250} \approx 0.77.$$

Note about the sampling

Note about the sampling: the data for these 8 tables were presumably obtained as "case-control samples." This means a fixed number of lung cancer cases were sampled from all lung cancer cases, and a fixed number of controls were sampled from all controls in the city.

In case/control sampling, the "cancer proportion" is not informative about the prevalence of lung cancer in any of these cities.

However the proportions of smokers among lung cancer cases (126/161 for Beijing) and among controls (100/161 for Beijing) are informative about the true proportions of smokers among cases and controls.