2.10: Cochran-Mantel-Haenszel Test

Learning Objectives

- To use the Cochran–Mantel–Haenszel test when you have data from 2×2 tables that you've repeated at different times or locations. It will tell you whether you have a consistent difference in proportions across the repeats.

When to use it

Use the Cochran–Mantel–Haenszel test (which is sometimes called the Mantel–Haenszel test) for repeated tests of independence. The most common situation is that you have multiple 2×2 tables of independence; you're analyzing the kind of experiment that you'd analyze with a test of independence, and you've done the experiment multiple times or at multiple locations. There are three nominal variables: the two variables of the 2×2 test of independence, and the third nominal variable that identifies the repeats (such as different times, different locations, or different studies). There are versions of the Cochran–Mantel–Haenszel test for any number of rows and columns in the individual tests of independence, but they're rarely used and I won't cover them.

Fig. 2.10.1 A pony wearing pink legwarmers

For example, let's say you've found several hundred pink knit polyester legwarmers that have been hidden in a warehouse since they went out of style in 1984. You decide to see whether they reduce the pain of ankle osteoarthritis by keeping the ankles warm. In the winter, you recruit 36 volunteers with ankle arthritis, randomly assign 20 to wear the legwarmers under their clothes at all times while the other 16 don't wear the legwarmers, then after a month you ask them whether their ankles are pain-free or not. With just the one set of people, you'd have two nominal variables (legwarmers vs. control, pain-free vs. pain), each with two values, so you'd analyze the data with Fisher's exact test.

However, let's say you repeat the experiment in the spring, with 50 new volunteers. Then in the summer you repeat the experiment again, with 28 new volunteers. You could just add all the data together and do Fisher's exact test on the 114 total people, but it would be better to keep each of the three experiments separate. Maybe legwarmers work in the winter but not in the summer, or maybe your first set of volunteers had worse arthritis than your second and third sets. In addition, pooling different studies together can show a "significant" difference in proportions when there isn't one, or even show the opposite of a true difference. This is known as Simpson's paradox. For these reasons, it's better to analyze repeated tests of independence using the Cochran-Mantel-Haenszel test.

Null hypothesis

The null hypothesis is that the relative proportions of one variable are independent of the other variable within the repeats; in other words, there is no consistent difference in proportions in the 2×2 tables. For our imaginary legwarmers experiment, the null hypothesis would be that the proportion of people feeling pain was the same for legwarmer-wearers and non-legwarmer wearers, after controlling for the time of year. The alternative hypothesis is that the proportion of people feeling pain was different for legwarmer and non-legwarmer wearers.

Technically, the null hypothesis of the Cochran–Mantel–Haenszel test is that the odds ratios within each repetition are equal to 1. The odds ratio is equal to 1 when the proportions are the same, and the odds ratio is different from 1 when the proportions are different from each other. I think proportions are easier to understand than odds ratios, so I'll put everything in terms of proportions. But if you're in a field such as epidemiology where this kind of analysis is common, you're probably going to have to think in terms of odds ratios.

How the test works

If you label the four numbers in a 2×2 test of independence like this:

$$\begin{matrix} a & b \\ c & d \end{matrix} \tag{2.10.1}$$

and

$$(a+b+c+d) = n \tag{2.10.2}$$

you can write the equation for the Cochran–Mantel–Haenszel test statistic like this:

$$X^2_{MH} = \frac{\{|\sum[a-(a+b)(a+c)/n]|-0.5\}^2}{\sum(a+b)(a+c)(b+d)(c+d)/(n^3-n^2)} \tag{2.10.3}$$

The numerator contains the absolute value of the difference between the observed value in one cell (a) and the expected value under the null hypothesis, $(a+b)(a+c)/n$, so the numerator is the squared sum of deviations between the observed and expected values. It doesn't matter how you arrange the 2×2 tables, any of the four values can be used as a. You subtract the 0.5 as a continuity correction. The denominator contains an estimate of the variance of the squared differences.

The test statistic, $X^2_{MH'}$, gets bigger as the differences between the observed and expected values get larger, or as the variance gets smaller (primarily due to the sample size getting bigger). It is chi-square distributed with one degree of freedom.

Different sources present the formula for the Cochran–Mantel–Haenszel test in different forms, but they are all algebraically equivalent. The formula I've shown here includes the continuity correction (subtracting 0.5 in the numerator), which should make the P value more accurate. Some programs do the Cochran–Mantel–Haenszel test without the continuity correction, so be sure to specify whether you used it when reporting your results.

Assumptions

In addition to testing the null hypothesis, the Cochran-Mantel-Haenszel test also produces an estimate of the common odds ratio, a way of summarizing how big the effect is when pooled across the different repeats of the experiment. This require assuming that the odds ratio is the same in the different repeats. You can test this assumption using the Breslow-Day test, which I'm not going to explain in detail; its null hypothesis is that the odds ratios are equal across the different repeats.

If some repeats have a big difference in proportion in one direction, and other repeats have a big difference in proportions but in the opposite direction, the Cochran-Mantel-Haenszel test may give a non-significant result. So when you get a non-significant Cochran-Mantel-Haenszel test, you should perform a test of independence on each 2×2 table separately and inspect the individual P values and the direction of difference to see whether something like this is going on. In our legwarmer example, if the proportion of people with ankle pain was much smaller for legwarmer-wearers in the winter, but much higher in the summer, and the Cochran-Mantel-Haenszel test gave a non-significant result, it would be erroneous to conclude that legwarmers had no effect. Instead, you could conclude that legwarmers had an effect, it just was different in the different seasons.

Examples

Example

When you look at the back of someone's head, the hair either whorls clockwise or counterclockwise. Lauterbach and Knight (1927) compared the proportion of clockwise whorls in right-handed and left-handed children. With just this one set of people, you'd have two nominal variables (right-handed vs. left-handed, clockwise vs. counterclockwise), each with two values, so you'd analyze the data with Fisher's exact test.

However, several other groups have done similar studies of hair whorl and handedness (McDonald 2011):

Study group	Handedness	Right	Left
white children	Clockwise	708	50
	Counterclockwise	169	13
	percent CCW	19.3%	20.6%

Study group	Handedness	Right	Left
British adults	Clockwise	136	24
	Counterclockwise	73	14
	percent CCW	34.9%	38.0%
Pennsylvania whites	Clockwise	106	32
	Counterclockwise	17	4
	percent CCW	13.8%	11.1%
Welsh men	Clockwise	109	22
	Counterclockwise	16	26
	percent CCW	12.8%	54.2%
German soldiers	Clockwise	801	102
	Counterclockwise	180	25
	percent CCW	18.3%	19.7%
German children	Clockwise	159	27
	Counterclockwise	18	13
	percent CCW	10.2%	32.5%
New York	Clockwise	151	51
	Counterclockwise	28	15
	percent CCW	15.6%	22.7%
American men	Clockwise	950	173
	Counterclockwise	218	33
	percent CCW	18.7%	16.0%

You could just add all the data together and do a test of independence on the 4463 total people, but it would be better to keep each of the 8 experiments separate. Some of the studies were done on children, while others were on adults; some were just men, while others were male and female; and the studies were done on people of different ethnic backgrounds. Pooling all these studies together might obscure important differences between them.

Analyzing the data using the Cochran-Mantel-Haenszel test, the result is $X^2_{MH}=6.07,\ 1 d.f.,\ P=0.014$. Overall, left-handed people have a significantly higher proportion of counterclockwise whorls than right-handed people.

Example

McDonald and Siebenaller (1989) surveyed allele frequencies at the *Lap* locus in the mussel *Mytilus trossulus* on the Oregon coast. At four estuaries, we collected mussels from inside the estuary and from a marine habitat outside the estuary. There were three common alleles and a couple of rare alleles; based on previous results, the biologically interesting question was whether the ***Lap***$^{\boldsymbol{94}}$ allele was less common inside estuaries, so we pooled all the other alleles into a "**non-*94***" class.

There are three nominal variables: allele (94 or non-94), habitat (marine or estuarine), and area (Tillamook, Yaquina, Alsea, or Umpqua). The null hypothesis is that at each area, there is no difference in the proportion of **Lap**$^{\mathbf{94}}$ alleles between the marine and estuarine habitats.

This table shows the number of 94 and non-94 alleles at each location. There is a smaller proportion of 94 alleles in the estuarine location of each estuary when compared with the marine location; we wanted to know whether this difference is significant.

Location	Allele	Marine	Estuarine
Tillamook	94	56	69
	non-94	40	77
	percent 94	58.3%	47.3%
Yaquina	94	61	257
	non-94	57	301
	percent 94	51.7%	46.1%
Alsea	94	73	65
	non-94	71	79
	percent 94	50.7%	45.1%
Umpqua	94	71	48
	non-94	55	48
	percent 94	56.3%	50.0%

The result is $X^2_{MH} = 5.05,\ 1 d.f.,\ P = 0.025$. We can reject the null hypothesis that the proportion of ***Lap*^{94}** alleles is the same in the marine and estuarine locations.

Example

Duggal et al. (2010) did a meta-analysis of placebo-controlled studies of niacin and heart disease. They found 5 studies that met their criteria and looked for coronary artery revascularization in patients given either niacin or placebo:

Study		Revascularization	No revasc.	Percent revasc.
FATS	Niacin	2	46	4.2%
	Placebo	11	41	21.2%
AFREGS	Niacin	4	67	5.6%
	Placebo	12	60	16.7%
ARBITER 2	Niacin	1	86	1.1%
	Placebo	4	76	5.0%
HATS	Niacin	1	37	2.6%
	Placebo	6	32	15.8%
CLAS 1	Niacin	2	92	2.1%
	Placebo	1	93	1.1%

There are three nominal variables: niacin vs. placebo, revascularization vs. no revascularization, and the name of the study. The null hypothesis is that the rate of revascularization is the same in patients given niacin or placebo. The different studies have different overall rates of revascularization, probably because they used different patient populations and looked for revascularization after different lengths of time, so it would be unwise to just add up the numbers and do a single 2×2 test. The result of the Cochran-Mantel-Haenszel test is $X^2_{MH} = 12.75,\ 1 d.f.,\ P = 0.00036$. Significantly fewer patients on niacin developed coronary artery revascularization.

Graphing the results

To graph the results of a Cochran–Mantel–Haenszel test, pick one of the two values of the nominal variable that you're observing and plot its proportions on a bar graph, using bars of two different patterns.

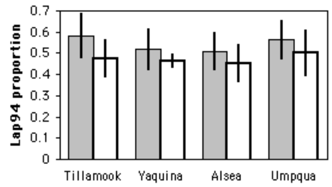

Fig. 2.10.2 **Lap^{94}** allele proportions (with 95% confidence intervals) in the mussel Mytilus trossulus at four bays in Oregon. Gray bars are marine samples and empty bars are estuarine samples.

Similar tests

Sometimes the Cochran–Mantel–Haenszel test is just called the Mantel–Haenszel test. This is confusing, as there is also a test for homogeneity of odds ratios called the Mantel–Haenszel test, and a Mantel–Haenszel test of independence for one 2×2 table. Mantel and Haenszel (1959) came up with a fairly minor modification of the basic idea of Cochran (1954), so it seems appropriate (and somewhat less confusing) to give Cochran credit in the name of this test.

If you have at least six 2×2 tables, and you're only interested in the *direction* of the differences in proportions, not the size of the differences, you could do a sign test.

The Cochran–Mantel–Haenszel test for nominal variables is analogous to a two-way anova or paired *t*–test for a measurement variable, or a Wilcoxon signed-rank test for rank data. In the arthritis-legwarmers example, if you measured ankle pain on a 10-point scale (a measurement variable) instead of categorizing it as pain/no pain, you'd analyze the data with a two-way anova.

How to do the test

Spreadsheet

I've written a spreadsheet to perform the Cochran–Mantel–Haenszel test cmh.xls. It handles up to 50 2×2 tables. It gives you the choice of using or not using the continuity correction; the results are probably a little more accurate with the continuity correction. It does not do the Breslow-Day test.

Web pages

I'm not aware of any web pages that will perform the Cochran–Mantel–Haenszel test.

R

Salvatore Mangiafico's *R Companion* has a sample R program for the Cochran-Mantel-Haenszel test, and also shows how to do the Breslow-Day test.

SAS

Here is a SAS program that uses PROC FREQ for a Cochran–Mantel–Haenszel test. It uses the mussel data from above. In the TABLES statement, the variable that labels the repeats must be listed first; in this case it is "location".

```
DATA lap;
INPUT location $ habitat $ allele $ count;
DATALINES;
Tillamook marine 94 56
Tillamook estuarine 94 69
Tillamook marine non-94 40
Tillamook estuarine non-94 77
Yaquina marine 94 61
Yaquina estuarine 94 257
Yaquina marine non-94 57
Yaquina estuarine non-94 301
```

```
Alsea marine 94 73
Alsea estuarine 94 65
Alsea marine non-94 71
Alsea estuarine non-94 79
Umpqua marine 94 71
Umpqua estuarine 94 48
Umpqua marine non-94 55
Umpqua estuarine non-94 48
;
PROC FREQ DATA=lap;
WEIGHT count / ZEROS;
TABLES location*habitat*allele / CMH;
RUN;
```

There is a lot of output, but the important part looks like this:

```
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic Alternative Hypothesis DF Value Prob
---------------------------------------------------------
1 Nonzero Correlation 1 5.3209 0.0211
2 Row Mean Scores Differ 1 5.3209 0.0211
3 General Association 1 5.3209 0.0211
```

For repeated 2×2 tables, the three statistics are identical; they are the Cochran–Mantel–Haenszel chi-square statistic, *without* the continuity correction. For repeated tables with more than two rows or columns, the "general association" statistic is used when the values of the different nominal variables do not have an order (you cannot arrange them from smallest to largest); you should use it unless you have a good reason to use one of the other statistics.

The results also include the Breslow-Day test of homogeneity of odds ratios:

```
Breslow-Day Test for
Homogeneity of the Odds Ratios
------------------------------
Chi-Square 0.5295
DF 3
Pr > ChiSq 0.9124
```

The Breslow-Day test for the example data shows no significant evidence for heterogeneity of odds ratios ($X^2 = 0.53$, $3 d.f.$, $P = 0.91$).

References

Picture of Flashprance the pony from Pony Paradise.

Cochran, W.G. 1954. Some methods for strengthening the common χ^2 tests. Biometrics 10: 417-451.

Duggal, J.K., M. Singh, N. Attri, P.P. Singh, N. Ahmed, S. Pahwa, J. Molnar, S. Singh, S. Khosla and R. Arora. 2010. Effect of niacin therapy on cardiovascular outcomes in patients with coronary artery disease. Journal of Cardiovascular Pharmacology and Therapeutics 15: 158-166.

Lauterbach, C.E., and J.B. Knight. 1927. Variation in whorl of the head hair. Journal of Heredity 18: 107-115.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22: 719-748.

McDonald, J.H. 2011. Myths of human genetics. Sparky House Press, Baltimore.

McDonald, J.H. and J.F. Siebenaller. 1989. Similar geographic variation at the *Lap* locus in the mussels *Mytilus trossulus* and *M. edulis*. Evolution 43: 228-231.