BIOS 6244 Analysis of Categorical Data
December 7, 2005 Lecture

Models for Matched Pairs (Chapter 9)

Categorical data in health sciences research often occur in *pairs* that are defined by the design of the study. For example, in a *cross-over design*, all the the study subjects receive both treatments A & B: ½ receive them in order AB & the other ½ receive them in order BA. Thus, each subject is "paired" with themselves and serves as their own control. The reason that both AB & BA are used in the study is because of the possibility of a carry-over effect; that is, patients who start on Drug A and later switch to Drug B may still be subject to the effects of Drug A.

Example (Exercise 9.7, p. 250)

9.7 A crossover experiment with 100 subjects compares two drugs used to treat migraine headaches. The response scale is success (+) or failure (−). Half the study subjects, randomly selected, used drug A the first time they get a migraine headache and drug B the next time. For them, 6 had responses $(A+, B+)$, 25 had responses $(A+, B-)$, 10 had responses $(A-, B+)$, and 9 had responses $(A-, B-)$. The other 50 subjects took the drugs in the reverse order. For them, 10 were $(A+, B+)$, 20 were $(A+, B-)$, 12 were $(A-, B+)$, and 8 were $(A-, B-)$.

Other study designs in which the data are paired included case-control studies, in which the controls (i.e., those without the disease) are matched on some objective criteria (e.g., age, gender) to the cases (i.e., those with the disease). An example of such a study is described on p. 232 of our textbook:

Table 9.3 ***Previous Diagnoses of Diabetes for Myocardial Infarction (MI)*** **Case-Control Pairs**

	MI Controls		
MI Cases	Diabetes	No Diabetes	Total
Diabetes	9	37	46
No Diabetes	16	82	98
Total	25	119	144

Source: J. L. Coulehan et al., *Amer. J. Public Health, 76:* 412–414 (1986). Reprinted with permission of the American Public Health Association. See also M. Pagano and K. Gauvreau, *Principles of Biostatistics* (Duxbury Press 1993, p. 319).

Table 9.3 illustrates results of a matched case-control study. A study of acute myocardial infarction (MI) among Navajo Indians matched 144 victims of MI according to age and gender with 144 individuals free of heart disease. Subjects were then asked whether they had ever been diagnosed as having diabetes ($x = 0$, no; $x = 1$, yes).

Note that in Table 9.3, the marginal totals give the # of "successes" in each group being compared: For the cases, 46/144 said they had previously been diagnosed with diabetes, whereas for the controls, 25/144 said they had.

Another commonly used design in pilot studies in clinical research is the "before" and "after" design (sometimes mistakenly called a *time series* by clinical researchers), in which study subjects are again "paired" with themselves and observed before *and* after receiving some treatment. (There typically is no control group.) For example, in a study of the effectiveness of botulism toxin ("Botox") as a treatment for acne, 7 adolescents with acne were examined before and after receiving Botox injections near the sites of their acne lesions.

(On pp. 226-227 of our text, Agresti describes an interesting example of another matched pairs design in which Canadian citizens were asked for their opinions about their Prime Minister in two surveys taken 6 months apart.)

In all of these examples, the subjects' responses are *paired* (i.e., correlated). In the cross-over design, we want to ultimately compare the subjects' responses under drugs A & B. However, Fisher's exact test cannot be used to make this comparison since the 2 groups receiving A & B are not independent – *they are the same subjects measured on 2 different occasions.* (Note that Fisher's exact test *could* be used to compare the subjects receiving AB with those receiving BA.)

In the case-control example, cases and controls cannot be treated as independent groups since correlation between the 2 groups has been deliberately introduced by matching on age and gender. [Note that if the cases and controls had been randomly selected from the population of all Navajo Indians, then the 2 groups *could* be treated as independent and Fisher's exact test could have been applied. That is why it is important to match *only* on known confounders (age and gender in the example). If the matching variables are not true confounders, then accounting for the matching in the analysis can adversely affect the performance of the statistical tests.]

In the Botox example, we wish to compare the severity of the acne lesions before and after administering Botox. Since the same adolescent is examined on two different occasions, the observations cannot be assumed to be independent, so, again the methods we have discussed previously for comparing 2 groups of subjects would not be appropriate.

Of course, the question now becomes "What *is* the appropriate method for analyzing paired categorical data?" (Sometimes this is called the problem of *dependent proportions*.) This is analogous to the problem of comparing 2 population means - if the groups can be assumed to be independent of each other, the *2-sample t-test* can be used. If the groups are not independent (i.e., paired in some way), then the *paired t-test* should be used. Examples include "before and after" studies, studies involving twins, and cross-over designs.

In this section, we will discuss methods for comparing 2 dependent or paired proportions that are analogous to the paired t-test for comparing 2 dependent or paired means.

Generally, Agresti will refer to the data as *matched pairs* even if no formal matching mechanism was used. A subject's response under 1 condition or on 1 occasion is "matched" to their response under another condition or on another occasion simply because the same subject is being observed twice.

The data in a matched pairs study are presented differently from what we have seen previously. In all of the examples we have considered thus far in which X & Y were both binary, the levels of the predictor (X) were represented by the rows and the levels of the outcome (Y) were represented by the columns in the 2x2 table. For matched pairs data, however, the same categories are used for both the rows and columns, with the rows representing one member of the pair and columns representing the other member of the pair. This special format for the 2x2 table is called the *canonical form* for matched pairs data.

For example, consider the case-control study involving Navajo Indians (Table 9.3).

Table 9.3 *Previous Diagnoses of Diabetes for Myocardial Infarction (MI)* Case-Control Pairs

	MI Controls		
MI Cases	Diabetes	No Diabetes	Total
Diabetes	9	37	46
No Diabetes	16	82	98
Total	25	119	144

Source: J. L. Coulehan et al., *Amer. J. Public Health, 76:* 412–414 (1986). Reprinted with permission of the American Public Health Association. See also M. Pagano and K. Gauvreau, *Principles of Biostatistics* (Duxbury Press 1993, p. 319).

Here, the rows refer to MI cases and the columns refer to MI controls. Both rows & columns are labelled Diabetes/No Diabetes. Thus, the main diagonal entries in the 2x2 table now indicate the # of cases and controls that "agreed" in terms of having the risk factor (i.e., Diabetes) or not. The off-diagonal entries in the 2x2 table now indicate the # of cases and controls that "disagreed" in terms of the risk factor.

Similarly, for the cross-over study described in Exercise 9.7, the data would be put in the following 2x2 table before proceeding with the matched pairs analysis:

		Treatment B: Success	Treatment B: Failure	
Treatment A	Success	16	45	61
	Failure	22	17	39
		38	62	100

Comparing Dependent Proportions (Sec. 9.1)

Let n_{ij} = # of subjects in the (i,j) cell of Table 9.3. Note that $n_{1+} = n_{11} + n_{12} = 46$ = # of patients with diabetes among those who suffered an MI (the cases) and $n_{+1} = n_{11} + n_{21} = 9 + 16 = 25$ = # of patients with diabetes among those who did not suffer an MI (the controls). Thus the sample proportions are $\frac{46}{144} = .32$ and $\frac{25}{144} = .17$. We need to compare these proportions in order to determine if diabetes is more likely among those who suffered an MI. However, these proportions are correlated due to the matching of the cases & controls and the statistical analysis much take this into account.

Let π_{ij} = true probability that a subject falls into cell (i,j) in Table 9.3. The true probabilities of having diabetes for the cases and controls are given by π_{1+} and π_{+1}, respectively. When $\pi_{1+} = \pi_{+1}$, we say that *marginal homogeneity* is present. Since $\pi_{1+} - \pi_{+1} = \pi_{11} + \pi_{12} - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}$, marginal homogeneity in a 2x2 table is equivalent to equality of the "off diagonal" probabilities, i.e., $\pi_{12} = \pi_{21}$. The table is then said to show *symmetry* across the main diagonal.

McNemar Test (Sec. 9.1.1)

The test of marginal homogeneity for matched (correlated) binary responses has null hypothesis H_0: $\pi_{1+} = \pi_{+1}$ or, equivalently, H_0: $\pi_{12} = \pi_{21}$. When H_0 is true, then we would expect to see about the same observed frequencies n_{12} and n_{21}. Let $n^* = n_{12} + n_{21}$ denote the total count in the 2 off-diagonal cells. Conditional on the value of n*, the allocation of the n* observations to one of the 2 off-diagonal cells is a binomial random variable (RV) with n* trials and probability of "success" π. Under the null hypothesis H_0: $\pi_{12} = \pi_{21}$, each of the n* observations has probability ½ of being in cell (1,2) and probability ½ of being in cell (2,1). So, n_{12} & n_{21} are the # of "successes" and "failures" for a binomial RV having n* trials and probability of success ½.

Thus, a test of H_0: $\pi_{12} = \pi_{21}$ can be performed using the binomial distribution to calculate the exact p-value. First, consider the one-sided alternative hypothesis H_a: $\pi_{1+} > \pi_{+1}$ or,

equivalently, H_a: $\pi_{12} > \pi_{21}$. From Table 9.3, $n_{12} = 37$, $n_{21} = 16$, and $n^* = 37 + 16 = 53$. The reference distribution (conditional on the value of n*) is a binomial with $n^* = 53$ & $\pi = .5$. The p-value for the 1-sided alternative above is then $\Pr(n_{12} \geq 37 \mid n^* = 53, \pi = .5) = .0027$ by SimCalc. For the 2-sided alternative H_a: $\pi_{1+} \neq \pi_{+1}$, the 2-tailed p-value would be 2(.0027) = .0054. Thus, there is very strong evidence of a positive association between diabetes & MI.

Since we know that any binomial distribution with $\pi = .5$ is symmetric, the normal approximation to the null distribution of n_{12} is quite good, even for n* as small as 10. Under this approximation, n_{12} has an approximate normal distribution with $\mu = ½\, n^*$ & $\sigma^2 = ¼\, n^*$. Thus, to use the normal approximation to the binomial to calculate the p-value, the test statistic would be

$$Z = \frac{n_{12} - \frac{n^*}{2}}{\sqrt{\frac{n^*}{4}}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

We know that $Z^2 \sim \chi^2(1)$. If we use

$$Z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}},\ df = 1$$

as the test statistic, the procedure is called *McNemar's Test.*

To apply McNemar's Test (which is always a 2-sided test) to the data in Table 9.3, we have

$$Z^2 = \frac{(37-16)^2}{37+16} = 8.32,\ df = 1$$

and p = .0039 by SimCalc. Compare this with the exact 2-tailed p-value of .0054 we obtained using the binomial.

Given the availability of modern statistical software that can easily calculate binomial probabilities even for large n, there is no reason to apply McNemar's Test as originally formulated. I recommend that you always apply the exact version using the binomial distribution.

Estimating Differences of Dependent Proportions (Sec. 9.1.2)

In addition to using the exact version of McNemar's method to test the equivalence of dependent proportions, we usually also want to find a CI for their true difference. The $\mathrm{Var}(\hat{\pi}_{1+} - \hat{\pi}_{+1})$ is

$$\frac{p_{1+}(1-p_{1+})+p_{+1}(1-p_{+1})-2(p_{11}p_{22}-p_{12}p_{21})}{n}. \qquad (31)$$

The 1st 2 terms in the numerator constitute the usual formula for the estimate of Var (p_{1+} - p_{+1}) when p_{1+} and p_{+1} are independent. The part in parentheses in the last term in the numerator is an estimate of Cov((p_{1+} , p_{+1}), which is needed as an"adjustment" for the fact that p_{1+} and p_{+1} are not independent.

Matched-pair data tend to exhibit a strong positive association (OR >> 1). This corresponds to $p_{11}p_{22} >> p_{12}p_{21}$, so there would be a considerable negative adjustment in the numerator of Equation (31) above. This would result in a smaller estimate of $\text{Var}(\hat{\pi}_{1+} - \hat{\pi}_{+1})$ and hence a larger z^2 test statistic and a smaller p-value. This illustrates one of the advantages of using matching (provided you have good matching criteria) or a repeated measures design (as in a cross-over design) – generally you will have greater power that if you had used independent samples of comparable size.

The square root of the estimated variance in Equation (31) can be used as the standard error in the derivation of an approximate 95% CI(π_{1+} - π_{+1}):

$$p_{1+} - p_{+1} \pm z_{\alpha/2}\text{SE}(p_{1+} - p_{+1}) = p_{1+} - p_{+1} \pm z_{\alpha/2}\sqrt{\frac{p_{1+}(1-p_{1+})+p_{2+}(1-p_{2+})-2(p_{11}p_{22}-p_{12}p_{21})}{n}}$$

=

$$(.3194-.1736) \pm 1.96\sqrt{\frac{.3194(1-.3194)+.1736(1-.1736)-2[(.0625)(.5694)-(.2569)(.1111)]}{144}}$$

= .1458 ± 1.96 (.0491) = .1458 ± .0962 = (.0496, .2420).

Thus, we can be 95% sure that the true difference in prevalence of diabetes between Navajo Indians who have suffered an MI and those who haven't is between .05 and .24.

StatXact can compute an exact 95% CI(π_{1+} - π_{+1}). For the data in Table 9.3, it is (.0378, .2443)

PROC FREQ in SAS can be used to perform the exact version of McNemar's test. The following SAS code will perform this test for the data in Table 9.3:

```
data navajo;
input case control count @@;
cards;
1 1 9   1 0 37
0 1 16  0 0 82
;
```

(SAS Code continued on next page.)

```
proc format;
value casefmt
      1='diab'
      0='no diab';

proc format;
value cntlfmt
      1='diab'
      0='no diab';

proc freq order=data; weight count;
   format case casefmt.;
   format control cntlfmt.;
   tables case*control / agree;
   exact  MCNEM;
   title 'Table 9.3';
   title2 'McNemar Test';
run;
```

The relevant SAS output is as follows:

Table 9.3
McNemar Test

The FREQ Procedure

Table of case by control

case / control	diab	no diab	Total
diab	9	37	46
	6.25	25.69	31.94
	19.57	80.43	
	36.00	31.09	
no diab	16	82	98
	11.11	56.94	68.06
	16.33	83.67	
	64.00	68.91	
Total	25	119	144
	17.36	82.64	100.00

Statistics for Table of case by control

McNemar's Test

Statistic (S)	8.3208
DF	1
Asymptotic Pr > S	0.0039
Exact Pr >= S	0.0055

Note that the above results produced by SAS agree with our hand calculations.

Extensions:

We can also compare k > 2 dependent proportions (e.g., the subjects in the Botox study were evaluated at baseline, 1, 2, 4, 8, and 12 weeks). The method to use with data of this type is called *Cochran's Q test.* The Cochran-Mantel-Haenszel approach can be used to adjust tests of dependent proportions for a stratified confounder. Logistic regression models can be easily adapted for use in matched case-control studies. Methods are also available for comparing dependent groups when there are more than 2 categories for the outcome variable (e.g., worse, same, improved). All of these methods are discussed in the remainder of Chapter 9 in our text.