

Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 15, part 3: McNemar's Test

Jinko Graham

2018-10-22

Paired Data and McNemar's Test

- ▶ The chi-square test is not appropriate when the study is designed to collect paired data.
- ▶ Example: Study of smoking habits over time.
 - ▶ A random sample of 2110 people were questioned about smoking status in 1980 and again in 1982.
 - ▶ Are smoking status and year associated? i.e, does the population proportion of smokers differ by year?
 - ▶ Test $H_0 : p_{1980} - p_{1982} = 0$ vs. $H_a : p_{1980} - p_{1982} \neq 0$, where p_{1980} and p_{1982} are the population proportions of smokers in 1980 and 1982, respectively.
- ▶ The data might look as follows:

person	Smoking	
	1980	1982
1	no	yes
2	no	no
3	yes	no
\vdots	\vdots	\vdots
2110	yes	yes

- ▶ The data could be arranged in a table as:

		Year		
		1980	1982	
Smoke	Yes	717	696	1413
	No	1393	1414	2807
		2110	2110	4220

- ▶ But this table is misleading: the 4220 observations that it reports are from 2110 double-counted people.
- ▶ The 4220 observations counted in the table are not independent, but rather paired observations from just 2110 people.

- ▶ To clarify the paired nature of the data, they are typically arranged as:

		1982 (after)		
		Smoker	Nonsmoker	
1980 (before)	Smoker	620	97	717
	Nonsmoker	76	1317	1393
		696	1414	2110

- ▶ The observations that are counted in the cells of this table are people and are independent.
- ▶ The cells with the same before- and after-status of the subject are called *concordant*.
- ▶ The cells with different before- and after-status are called *discordant*.
- ▶ McNemar's test is a contrast between the discordant cells, and ignores the concordant cells.

		1982 (after)		
		Smoker	Nonsmoker	
1980 (before)	Smoker	620	$r = 97$	717
	Nonsmoker	$s = 76$	1317	1393
		696	1414	$n = 2110$

- ▶ Our sample of 2110 individuals has 620 continuing smokers, $r = 97$ quitters and $s = 76$ starters in 1982.
- ▶ In 1980 and 1982, the sample proportions of smokers are, respectively,
 - ▶ $\hat{p}_{1980} = (620 + r)/n = (620 + 97)/2110$
 - ▶ $\hat{p}_{1982} = (620 + s)/n = (620 + 76)/2110$
- ▶ So, $\hat{p}_{1980} \neq \hat{p}_{1982}$ when $r \neq s$; or when $\hat{p}_{start} \neq \hat{p}_{quit}$, where
 - ▶ $\hat{p}_{start} = s/n$ is the proportion starting in 1982
 - ▶ $\hat{p}_{quit} = r/n$ is the proportion quitting in 1982.
- ▶ Re-express $H_0 : p_{1980} - p_{1982} = 0$ vs. $H_a : p_{1980} - p_{1982} \neq 0$ as
 - ▶ $H'_0 : p_{quit} = p_{start}$ vs. $H'_a : p_{quit} \neq p_{start}$.

McNemar's Test

- ▶ Base the hypothesis test on the difference in observed proportions $\hat{p}_{start} - \hat{p}_{quit} = (r - s)/n$ and its standard error.
- ▶ We'll skip the derivation but the test statistic ends up being:

$$X^2 = \frac{(r - s)^2}{r + s} \sim \chi_1^2.$$

- ▶ If the number of quitters, r , is very different from the number of starters, s , the statistic X^2 is **big** and we reject H_0 in favour of H_a .
- ▶ An alternate form that uses a continuity correction for small samples (text, page 351) is

$$X^2 = \frac{(|r - s| - 1)^2}{r + s} \sim \chi_1^2.$$

McNemar's Test for the Smoking Data

- ▶ We have $r = 97$ and $s = 76$. The test statistic with continuity correction is

$$\frac{(|97 - 76| - 1)^2}{97 + 76} = 2.31$$

and the corresponding p-value is 0.128 (see R demo).

- ▶ Taking $\alpha = .05$, there is insufficient statistical evidence to conclude that smoking status is associated with year (the pvalue 0.128 is > 0.05).

Other Examples of Paired Data

- ▶ Scoring individuals from the same matched pair.
 - ▶ e.g. case-control pairs in which the control has been matched to the case on a number of characteristics.
- ▶ Scoring the same experimental unit with two different techniques
- ▶ Ratings of the same experimental unit by two different raters
- ▶ Genetic variants that were transmitted/not transmitted from a parent (experimental unit) to its offspring

Example: Transmission/Disequilibrium Test (TDT)

- ▶ Spielman *et al.*, 1993 click developed a test for genetic linkage that is an application of McNemar's test
- ▶ DNA segments that are physically close together on a chromosome, or “genetically linked”, tend to be co-transmitted from parent to offspring.
 - ▶ A DNA marker that is genetically linked to a disease-causing mutation (i.e. they are physically close on the chromosome) should tend to be co-transmitted with the disease.
- ▶ Application to type 1 autoimmune diabetes (T1D):
 - ▶ A genetic marker called 5'FP located near the insulin gene was associated with T1D in the population. Is 5'FP linked to a disease-causing mutation?
 - ▶ If so, we should see certain variants of 5'FP over-represented in transmissions from parents to children affected by T1D.

Diabetes Data from Spielman *et al.*

- ▶ The DNA marker had two variants denoted 1 and X.
- ▶ Study of 124 parents of children with T1D
 - ▶ All parents selected to carry a 1 and an X at the DNA marker.
- ▶ The dataset has
 - ▶ one row for each parent, and
 - ▶ two columns, one for the variant that was transmitted from the parent to the affected child, and one for the variant that was not transmitted.

##	transmitted	untransmitted
## 1	1	X
## 2	1	X
## 3	1	X
## 4	1	X
## 5	1	X
## 6	1	X
## 7	1	X
## 8	1	X

McNemar's test on T1D Data

- ▶ Cross tabulating transmitted and untransmitted variables for the heterozygous parents yields the following table.

```
##           untransmitted
## transmitted  1  X
##           1  0 78
##           X 46  0
```

- ▶ McNemar's test applied to the T1D data (see R demo):

```
##
## McNemar's Chi-squared test
##
## data:  tt
## McNemar's chi-squared = 8.2581, df = 1, p-value = 0.004057
```

- ▶ There is strong evidence that the 1 variant is preferentially transmitted over the X variant to the affected child.
 - ▶ The marker near the insulin gene is genetically linked to T1D.