

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

## Chapter 15, part 3: McNemar's Test (take 2)

Jinko Graham

2018-10-24

## 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 research question is about the association between smoking status and year. So we might arrange the data 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.
- ▶ If we were to apply a chi-square test to this table, we'd be testing whether a person's smoking statuses in 1982 and 1980 are associated.

- ▶ The chi-square test addresses the question: “Is a person’s 1982 smoking status independent of their 1980 smoking status?”, or

$$H_0 : P(1982 \text{ smoker} | 1980 \text{ smoker}) = P(1982 \text{ smoker} | 1980 \text{ non-smoker})$$

vs.

$$H_a : P(1982 \text{ smoker} | 1980 \text{ smoker}) \neq P(1982 \text{ smoker} | 1980 \text{ non-smoker})$$

- ▶ **Not** the research question we’re interested in.
- ▶ Want to test whether or not the population proportion of smokers in 1980,  $p_{1980}$ , is the same as the population proportion of smokers in 1982,  $p_{1982}$ ; i.e., test

$$H_0 : p_{1982} = p_{1980} \text{ vs. } H_a : p_{1982} \neq p_{1980}.$$

## How to use the table to test our hypotheses?

		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 of

$$H'_0 : p_{quit} = p_{start} \text{ vs. } H'_a : p_{quit} \neq p_{start}$$

on the difference in observed proportions

$$\hat{p}_{start} - \hat{p}_{quit} = (r - s)/n$$

and its standard error.

- ▶ Skip the derivation but the test statistic ends up being:

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

- ▶ If the number of quitters,  $r$ , is very different from the number of starters,  $s$ , the statistic  $\chi^2$  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

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

## 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  $> 0.05$ ).



# Notes

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

- ▶ In the smoking-example table,
  - ▶ Cells with the same before- and after-status of the subject are called *concordant*.
  - ▶ Cells with different before- and after-status are called *discordant*.
- ▶ In general, cells that are diagonal entries are *concordant* and cells that are off-diagonal entries are *discordant*.
- ▶ Note that McNemar's test is a contrast between the discordant cells only, and ignores the concordant cells.

# 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
- ▶ Scoring genetic variants from the same parent for transmission and non-transmission to an offspring.

## Example: Transmission/Disequilibrium Test (TDT)

- ▶ Spielman *et al.*, 1993 click applied McNemar's test to a problem in medical genetics.
- ▶ 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 physically close to a disease-causing mutation tends to be co-transmitted with the disease.
- ▶ Application to autoimmune or type 1 diabetes (T1D):
  - ▶ Is the DNA marker 5'FP (near the insulin gene) *linked* to a disease-causing mutation?
  - ▶ If so, certain variants of 5'FP will be over-represented in transmissions from parents to children affected by T1D.

## Diabetes Data from Spielman *et al.*

- ▶ The DNA marker had two variants, “1” and “X”.
- ▶ Study of 124 parents of children with T1D
  - ▶ Parents chosen to carry both a 1 and an X at the DNA marker.
- ▶ Is variant type associated with transmission status?
  - ▶ e.g., test  $H_0 : P(1|\text{transmitted}) = P(1|\text{untransmitted})$   
vs.  $H_a : P(1|\text{transmitted}) \neq P(1|\text{untransmitted})$
- ▶ The dataset has a 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

- ▶ The research question is about whether variant type is associated with transmission status and so we might arrange the data in a table as:

		Transmitted		
		yes	no	
Variant X	1	78	46	124
		46	78	124
		124	124	248

- ▶ But this table is misleading: the 248 observations that it reports are the outcomes of transmission events from 124 double-counted parents.
- ▶ The 248 observations counted in the table are not independent, but rather paired observations from just 124 parents.

- ▶ To clarify the paired nature of the data, cross-tabulate the transmitted and untransmitted variables in the original dataset.

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

- ▶ The observations that are counted in the cells of this table are parents and are independent.
- ▶ If we were to apply a chi-square test to this table, we'd be testing whether a (1,X) parent's transmitted variant is associated with his/her untransmitted variant.
  - ▶ We don't need a test to see immediately that they are perfectly negatively dependent, *by definition*.
  - ▶ If one variant gets transmitted then the other one does not
- ▶ A chi-square test is not relevant to the research question.

# McNemar's test on T1D Data

- ▶ Research question: Is variant type associated with transmission status?
  - ▶ i.e.,  $H_0 : P(1|\text{transmitted}) = P(1|\text{untransmitted})$   
vs.  $H_a : P(1|\text{transmitted}) \neq P(1|\text{untransmitted})$ .
- ▶ Use McNemar's test (see R Demo):

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

- ▶ Strong evidence that variant type is associated with transmission status.

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

- ▶ In particular, the “1” variant appears to be preferentially transmitted over the “X” variant to the affected child.
- ▶ Can conclude that the DNA marker 5'FP is genetically linked to T1D.
- ▶ 5'FP is a DNA marker on chromosome 11, very close to the insulin gene.
  - ▶ Makes biological sense that DNA variation around the insulin gene would affect the risk of type 1 diabetes.
  - ▶ Body attacks and kills all the insulin-producing cells in the pancreas.