MSH4

Fundamentals of Statistical Consulting

Week 9 **Association coefficients for binary data**

Jean Yang and Jennifer Chan

Example: the matching of absence or presence of a dominant genetic marker

```
(x, y)
1 1
1 1
1 1
1 0
1 1
0 1
1 1
1 1
1 1
1 1
```

Why the association is NOT significant given a high proportion of agreement?

		Y		
		1	0	
X	1	a	b	$n_{1}. = np_{1}. = a + b$
	0	c	d	$n_{0}. = np_{0}. = c + d$
		$n._{1} = np._{1} = a + c$	$n._{0} = np._{0} = b + d$	n

Measures that include *d*

For nominal variables that are mutually exclusive, e.g. true or false, male or female, a and d should be equally weighted.

Yule, 1912: $$s_{\mathrm{phi}} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Equivalent to Pearson’s correlation applied to binary data and the numerator is covariance.

Sokal and Michener, 1958: $$s_{\mathrm{SM}} = \frac{a+d}{a+b+c+d}$$

Rogers and Tanimoto, 1960: $$s_{\mathrm{RT}} = \frac{a+d}{a+2(b+c)+d}$$

Cohen, 1960: $$s_{\mathrm{Cohen}} = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$$

Sokal and Sneath, 1963: $$s_{\mathrm{SS2}} = \frac{2(a+d)}{b+c+2(a+d)}$$

Sokal and Sneath, 1963: $$s_{\mathrm{SS3}} = \frac{1}{2n}\left(\frac{a}{a+d} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d}\right)$$

Sokal and Sneath, 1963: $$s_{\mathrm{SS4}} = \frac{ad}{\sqrt{(a+d)(a+c)(c+d)(b+d)}}$$

Measures that do not include d

However in some cases, the negative match d may dominate and should not contribute to similarity.

Similarity measures:

Jaccard, 1912: $s_{\text{Jac}} = \frac{a}{a+b+c}$ (no. of shared 1 to total no that contain 1)

Gleason, 1920: $s_{\text{Gleas}} = \frac{2a}{2a+b+c}$ (twice the wt, a few match relative to mismatch)

Kulczynski, 1927: $s_{\text{Kul}} = \frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$

Driver and Kroeber, 1932: $s_{\text{DK}} = \frac{a}{\sqrt{(a+b)(a+c)}}$

Sokal and Sneath, 1963: $s_{\text{SS1}} = \frac{a}{a+2(b+c)}$

Dissimilarity measures are defined the other way round.

Properties

Let $S(x,y)$ be the similarity coefficient between x and y.

Basic: $S(x,x) \geq S(x,y)$ and $S(y,y) \geq S(x,y)$.

Symmetric: $S(x,y) = S(y,x)$.

E.g. $S_{\text{phi}} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$ and $S_{\text{Jac}} = \frac{a}{a+b+c}$

Complete: $S(x,x) = 1$ E.g. $S_{\text{phi}}, S_{\text{Jac}}$

Independence: odds ratio (OR)$= \frac{a/b}{c/d} = \frac{ad}{bc} \in (0,\infty)$.

If OR $= 1$, equal likely in 0 and 1 groups. If OR > 1, more likely in the first group.

Transformed OR to within (-1,1):

Yule, 1900: $S_{\text{Yule1}} = \frac{ad-bc}{ad+bc}$

Yule, 1912: $S_{\text{Yule2}} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$

Example

(x, y)
1 1
1 1
1 1
1 0
1 1
0 1
1 1
1 1
1 1
1 1

		Y		
		1	0	
X	1	8	1	9
	0	1	0	1
		9	1	10

Different coefficients give the following results:

Use d:

Phi=(ad-bc)/ √(a+b)(a+c)(b+d)(c+d)=(8x0-1x1)/ √(9x1x9x1)= **-0.11**

SM= (a+d)/(a+b+c+d) = 8/(8+1+1+0)= **0.8**

RT=(a+d)/(a+2(b+c)+d) = 8/(8+2+2+0)= **0.67**

Cohen=2(ad-bc)/((a+b)(b+d)+(a+c)(b+d))=2(8x0-1x1)/(9x1+9x1)= **-0.11**

SS2=2(a+d)/(2a+b+c+2d)=2(8+0)/(2x8+1+1+2x0)= **0.89**

SS3=(a/(a+b)+a/(a+c)+d/(c+d)+d/(b+d))/4=(8/9+8/9+0/1+0/1)/4= **0.44**

SS4=ad/√(a+b)(a+c)(b+d)(c+d)= **0**

Not use d:

Jac=a/(a+b+c)=8/(8+1+1)= **0.8**

Gleas=2a/(2a+b+c)=2x8/(2x8+1+1)= **0.89**

Kul=(a/(a+b)+a/(a+c))/2=(8/9+8/9)/2= **0.89**

DK=8/$\sqrt{9 \times 9}$= **0.89**

OR and its transformation

OR=ad/bc=8x0/1x1= **0**

Yule1=(ad-bc)/(ad+bc)=(0-1)/(0+1)= **-1**

Yule2=($\sqrt{ad} - \sqrt{bc}/(\sqrt{ad} + \sqrt{bc}$)= **-1**

Values differ widely. No test of significant!

Test for independence

Pearson's chi-squared test for independence (larger sample size)

Same as between-subjects z-test for 2 proportions $z = \dfrac{\frac{n_{11}}{n_{1\cdot}} - \frac{n_{01}}{n_{0\cdot}}}{\sqrt{\frac{n_{\cdot 1}}{n} \times \frac{n_{\cdot 0}}{n} \left(\frac{1}{n_{1\cdot}} + \frac{1}{n_{0\cdot}} \right)}}$ (1)

Consider a 2x2 contingency table. Under H_0: $p_{ij} = p_{i\cdot}p_{\cdot j}$,

		Y		
		1	0	
X	1	$p_{1\cdot}p_{\cdot 1}$	$p_{1\cdot}p_{\cdot 0}$	$p_{1\cdot}$
	0	$p_{0\cdot}p_{\cdot 1}$	$p_{0\cdot}p_{\cdot 0}$	$p_{0\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 0}$	1

the χ^2 statistic for the 2x2 contingency table using (1) is

$$\chi_0^2 = \frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)} \sim \chi_1^2$$

where each cell has expected count of at least 5. Note the relationship:

$$s_{\text{phi}} = \sqrt{\frac{\chi^2}{n}} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Yates' correction for continuity: $\chi^2_{\text{Yates}} = \dfrac{n(|ad-bc|-n/2)^2}{(a+b)(a+c)(b+d)(c+d)}$ but mostly not needed.

Fisher exact test for independent binary data (smaller sample size)

Let the cell frequencies be represented by a, b, c, d, and the marginal totals represented by a+b, c+d, a+c, b+d, and n.

	1	**0**	**Totals**
1	a	b	$\boldsymbol{a+b}$
0	c	d	$\boldsymbol{c+d}$
Totals	$a+c$	$b+d$	n

If there are no systematic association between the variables X and Y, the probability of cell frequencies, a, b, c, d, given fixed marginal totals $\boldsymbol{a+b}$, $\boldsymbol{c+d}$, etc., are given by the hypergeometric rule:

$$\frac{\frac{(a+b)!}{a!b!}+\frac{(c+d)!}{c!d!}}{\frac{n!}{(a+c)!(b+d)!}} \text{ which is the same as } \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

The degree of association between variables X and Y can be measured by the absolute difference

$$\left|\frac{a}{a+b}-\frac{c}{c+d}\right|$$

We calculate the probability of that particular array plus the probabilities of all other possible arrays whose degree of disproportion is equal to or greater than that of the observed array. Thus, for the observed array

2	7	**9**
8	2	**10**
10	9	19

the one-tailed probability would be the sum of the separate probabilities for the arrays

		probability
2	7	
8	2	0.01754
1	8	
9	1	0.00097
0	9	
10	0	0.00001

sum = 0.01852 (one-tailed probability)

And the two-tailed probability would be that sum plus the sum of the separate probabilities for the arrays of equal or greater disproportion at the other extreme:

		probability
8	1	
2	8	0.00438
9	0	
1	9	0.00011

sum = 0.00449

The two-tailed probability = 0.01852 + 0.00449 = 0.02301

McNemar test for paired dependent binary data:

Same as within-subjects z-test for equality of 2 proportions.

Test if the before and after (from same person) marginal proportions are equal, i.e. $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$ implies $p_b = p_c$.

Condition on b and c only, the hypotheses are $H_0 : p_b = p_c = 0.5$ vs $H_1 : p_b \neq p_c$

where $p_b = p_c$ are conditional probabilities. Under H_0,

Observed	Expected	O-E	$(O\text{-}E)^2/E$
b	0.5(b+c)	0.5(b-c)	$0.25(b\text{-}c)^2/(0.5(b+c))$
c	0.5(b+c)	0.5(c-b)	$0.25(b\text{-}c)^2/(0.5(b+c))$
b+c	b+c	0	$(b\text{-}c)^2/(b+c)$

The test statistic is

$$\chi^2_{\mathrm{MN}} = \frac{(b-c)^2}{(b+c)} \sim \chi^2_1$$

Note a and d do not contribute to the decision though the number $a + b + c + d$ can be large. The exact test is the binomial sign test and Liddell’s exact test.

Difference between Chi-square test and McNemar test

1. The subjects are tested for infections from X at different times. Want to know if *the proportions of positive for X after is related to the proportion of positive for X before:*

```
             After
            |no  |yes|
Before|No   |1157|35 |
      |Yes  |220 |13 |

results of chi-squared test:
Chi^2 =  4.183     d.f. =  1     p =  0.04082

results of McNemar's test:
Chi^2 =  134.2     d.f. =  1     p =  4.901e-31
```

2. Instead of before and after, measure two different infections, X & Y at one time point (Before → X; After → Y).
Does higher proportions of one infections relate to higher proportions of Y"

Which test?

Q1: Chi-squared test assesses whether Before and After are independent. That is, are people who were sick beforehand more likely to be sick afterwards than people who have never been sick. Instead of whether Before and After are independent, one certainly wants to know if the treatment works (a question chi-squared does not answer). Specifically, one wants to run a within-subjects **z**-test of equality of proportions. That is what McNemar's test is.

		After		
		No	Yes	
Before	No	1157	**35**	1192
	Yes	**220**	13	233
		1377	48	1425

The proportion of yes before $= \frac{220+13}{1425}$ The proportion of yes after $= \frac{35+13}{1425}$

The 13 observations of yes to both before and after add no distinct information about the change in the proportion of yes. The only distinct information about the before and after proportions of yes is the numbers 220 and 35.
This is a binomial signtest of 220/(220+35) against a null proportion of 0.5.

```
mat = as.table(rbind(c(1157,35),c(220,13)))
colnames(mat) <- rownames(mat) <- c("No", "Yes")
```

```
names(dimnames(mat)) = c("Before", "After")
mat
#       After
# Before    No  Yes
#     No  1157   35
#     Yes  220   13
#
mcnemar.test(mat, correct=FALSE)
#  McNemar's Chi-squared test
#
# data:  mat
# McNemar's chi-squared = 134.2157, df = 1, p-value < 2.2e-16

binom.test(c(220, 35), p=0.5)  #exact one proportion sign test
#  Exact binomial test
#
# data:  c(220, 35)
# number of successes = 220, number of trials = 255, p-value < 2.2e-16
# alternative hypothesis: true probability of success is not equal to
0.5
# 95 percent confidence interval:
#  0.8143138 0.9024996
# sample estimates:
# probability of success
#              0.8627451
```

Q2. If we didn't take the within-subjects nature into account, one would have a slightly less powerful test of the equality of two proportions:

		Y		
		No	Yes	
X	No	1157	35	1192
	Yes	220	13	233
		1377	48	1425

The proportion of yes to X $= \dfrac{220+13}{1425}$ The proportion of yes to Y $= \dfrac{35+13}{1425}$

But the 13 observations of yes to both X and Y should be included. Hence information for the two proportions are (233, 1192) and (48,1377) which are the two marginals. Their sum is 2850, twice of 1425!

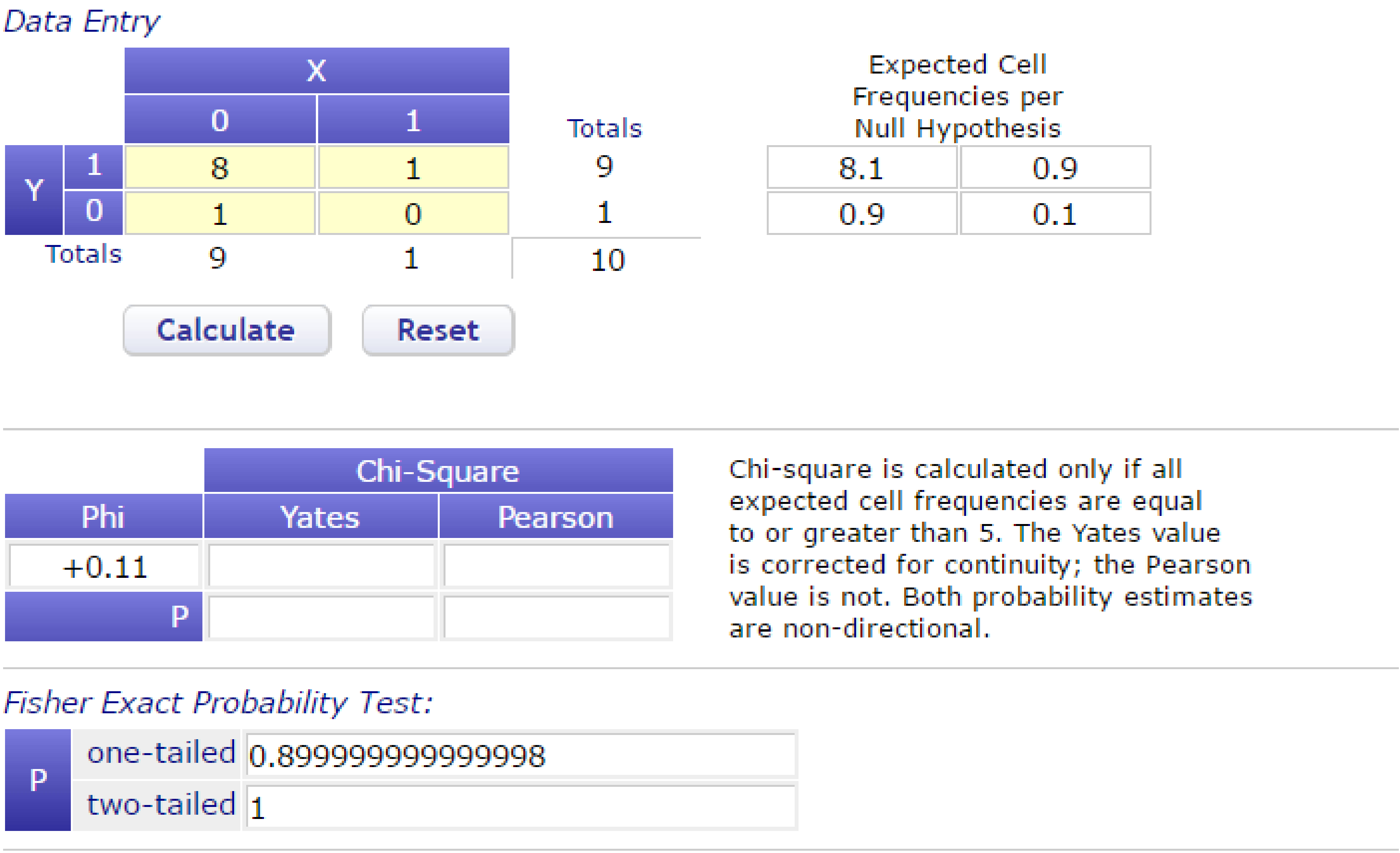
```
matm=as.table(rbind(margin.table(mat, 1),margin.table(mat, 2)))
colnames(matm) <- rownames(matm) <- c("No", "Yes")
names(dimnames(matm)) = c("X", "Y")
matm
#      Y
# X       No  Yes
#   No  1192  233
#   Yes 1377   48
#
```

```
chisq.test(matm,correct = F)
#
#         Pearson's Chi-squared test
#
# data:  matm
# X-squared = 135.1195, df = 1, p-value < 2.2e-16
#
prop.test(rbind(margin.table(mat, 1), margin.table(mat, 2)),
correct=FALSE)
#
#  2-sample test for equality of proportions without continuity
#  correction
#
# data:  rbind(margin.table(mat, 1), margin.table(mat, 2))
# X-squared = 135.1195, df = 1, p-value < 2.2e-16
# alternative hypothesis: two.sided
# 95 percent confidence interval:
#  0.1084598 0.1511894
# sample estimates:
#    prop 1    prop 2
# 0.9663158 0.8364912
```

The difference in chi-square of 135.1195 from 134.2157 is small relative due to the minor overlap of 13 but the sample size here is double.

Calculator

http://vassarstats.net/tab2x2.html



Data Entry

		X		Totals
		0	1	
Y	1	8	1	9
	0	1	0	1
Totals		9	1	10

Expected Cell Frequencies per Null Hypothesis

8.1	0.9
0.9	0.1

	Chi-Square	
Phi	Yates	Pearson
+0.11		
P		

Chi-square is calculated only if all expected cell frequencies are equal to or greater than 5. The Yates value is corrected for continuity; the Pearson value is not. Both probability estimates are non-directional.

Fisher Exact Probability Test:

P	one-tailed	0.899999999999998
	two-tailed	1

```
> mat = as.table(rbind(c(8, 1), c(1, 0) ))
> colnames(mat) <- rownames(mat) <- c("Yes", "No")
> names(dimnames(mat)) = c("Before", "After")
> mat
      After
Before Yes No
   Yes   8  1
   No    1  0
> margin.table(mat, 1)
Before
Yes  No
  9   1
> margin.table(mat, 2)
After
Yes  No
  9   1
> sum(mat)
[1] 10
> mcnemar.test(mat, correct=FALSE)

        McNemar's Chi-squared test
```

```
data:  mat
McNemar's chi-squared = 0, df = 1, p-value = 1

> binom.test(c(1, 1), p=0.5)

        Exact binomial test

data:  c(1, 1)
number of successes = 1, number of trials = 2, p-value = 1
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.01257912   0.98742088
sample estimates:
probability of success
                   0.5
>
> prop.test(rbind(margin.table(mat, 1), margin.table(mat, 2)), correct=FALSE)  #between subject

        2-sample test for equality of proportions without continuity correction
```

```
data:  rbind(margin.table(mat, 1), margin.table(mat, 2))
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2629568   0.2629568
sample estimates:
prop 1 prop 2
   0.9    0.9

Warning message:
In prop.test(rbind(margin.table(mat, 1), margin.table(mat, 2)),  :
  Chi-squared approximation may be incorrect
```

All show insignificant result!