Contigency Tables, Two Proportions Test, and McNemar's Test

Contingency Tables

For two categorical variables X with I categories, and Y with J categories, classification of subjects on both variables have IJ possible combinations. Typically one of them, say, Y is a response variable and X is an explanatory variable, we are interested in the `conditional distribution` of Y given the categories of X.

A rectangular table, having I rows for categories of X and J columns for categories of Y, with frequencies in the (IJ) cells is called a contingency table (Karl Pearson, 1904) or a cross-classification table. A contingency table with I rows and J columns is called an $I \times J$ table.

For example, the 2×2 contingency table, summarizes the two categorical variables, say, X, Y both having two categories. In particular, categorizing patients by their favorable or unfavorable response to two different drugs.

Chi-squared test

Suppose we are interested in testing the hypothesis that there is association between two categorical/nominal variables arranged as a contingency table. We can use what is called the chi-squared test.

Example - 1

Consider the 2×2 table below which displays the results of a study which investigated the effectiveness of bicycle safety helmets in preventing head injury.

	Wearing Helmet		
Head Injury	Yes	No	Total
Yes	17	218	235
No	130	428	558
Total	147	646	793

To examine the effectiveness of bicycle safety helmets, we can investigate whether there is an association between the incidence of head injury and the use of helmets among individuals who have been involved in accidents. A plausible test for this can be stated as follows:

$$H_0 : p_1 = p_2, \quad \text{versus} \quad H_a : p_1 \neq p_2,$$

p_1 is the proportion of persons sustaining head injuries among the population of individuals wearing a safety helmet at the time of the accident, and p_2 is the proportion of persons sustaining head injuries among those not wearing a helmet. Note that these two populations can be treated as independent.

Or equivalently,

H_0 : There is no association between `head injury` and `helmet wearing`

versus

H_a: There is association between `head injury` and `helmet wearing`

We can proceed with the two proportion test as before, but another approach would be to calculate the expected frequencies under the null hypothesis of equal proportions.

Note we have four observed frequencies $(O_i, i = 1, 2, 3, 4)$ given in the table as 17, 218, 130, 428. Using the binomial distribution properties we can calculate the corresponding expected frequencies $(E_i, i = 1, 2, 3, 4)$ under the null hypothesis. Then we calculate the test statistic as follows:

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

The probability distribution of the test statistic, χ^2 has a `Chi-squared` distribution with $(2-1)(2-1) = 1$ degrees of freedom.

The test of association between a row variable and a column variable, where both are arranged in a $r \times c$ contingency table is called a `Chi-squared` test.

In general, the test statistic

$$\chi^2 = \sum_{i=1}^{rc} \frac{(O_i - E_i)^2}{E_i}$$

for the test of association between a row variable and a column variable, where both are arranged in a $r \times c$ contingency table has a `Chi-squared` distribution with $(r-1)(c-1)$ degrees of freedom.

```
HI <- as.table(rbind(c(17,218),c(130,428)))
dimnames(HI) <- list(
  H_inj = c("Yes", "No"),
  Helmet = c("Yes", "No")
)
HI
```

```
##      Helmet
## H_inj Yes  No
##   Yes  17 218
##   No  130 428
```

```
chisq.test(HI)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  HI
## X-squared = 27.202, df = 1, p-value = 1.833e-07
```

The observed value of the `chi-squared` test statistic is 27.202. Note the test statistic has a chi-squared distribution with 1 degree of freedom.

The p-value is close to zero so we reject the null hypothesis that the proportion of persons sustaining head injuries among the population of individuals wearing a safety helmet at the time of the accident and the proportion of persons sustaining head injuries among those not wearing a helmet are the same.

In other words, we reject the hypothesis that there is no association between `head injury` and `helmet wearing` and conclude that there is association between `head injury` and `helmet wearing`.

McNemar's Test

Note that in the previous example, responses from two independent populations were cross tabulated in a 2×2 contingency table. In many applications, a 2×2 table contains information collected from dependent populations or matched pairs.

For example, in case-control studies, cases are often matched to controls on the basis of demographic characteristics. Interest lies in determining whether there is a difference between control exposure to a risk factor and case exposure to the same risk factor.

Other examples of matched pairs are left eye and right measurements, and husband and wife voting preferences.

The general layout of data from a study on matched pairs can be represented as follows:

	Response I		
Response II	Yes	No	Total
Yes	n_{11}	n_{12}	$n_{1.}$
No	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

Typically we are interested to know whether

$$p_1 = \frac{n_{.1}}{n}$$

and

$$p_2 = \frac{n_{1.}}{n}$$

are the same. Note these two proportions are not independent. The null and alternative hypotheses are written as follows:

H_0:There is no association between the responses I and II.

or

$$p_1 = p_2.$$

versus the alternative,

H_a: There is association between the responses I and II.

or

$$p_1 \neq p_2.$$

The test statistic is a chi-squared statistic

$$\chi^2_M = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})}$$

that has a chi-squared distribution with 1 degrees of freedom.

Myocardial Infarction and Diabetes

Consider a study that investigates acute myocardial infarction among members of the Navajo Nation. In the study, 144 victims of acute myocardial infarction were age- and sex-matched with 144 individuals free of heart disease. The members of each pair were then asked whether they had ever been diagnosed with diabetes. The responses are cross tabulated as follows:

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

Here the null hypothesis of interest is

H_0:There is no association between diabetes and the occurrence of acute myocardial infarction.

or the proportion of diabetes with MI and the proportion of diabetes with No MI are the same.

versus the alternative,

H_a:There is association between diabetes and the occurrence of acute myocardial infarction.

or the proportion of diabetes with MI and the proportion of diabetes with No MI are not the same.

```
MI <- as.table(rbind(c(9,37),c(16,82)))
dimnames(MI) <- list(
  "MI" = c("diabetes", "no diabetes"),
  "No MI" = c("diabetes", "no diabetes")
)
MI
```

```
##              No MI
## MI            diabetes no diabetes
##   diabetes           9          37
##   no diabetes       16          82
```

```
mcnemar.test(MI)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  MI
## McNemar's chi-squared = 7.5472, df = 1, p-value = 0.00601
```

The observed value of the `chi-squared` test statistic is 7.55 with p-value 0.006. Note the test statistic has a chi-squared distribution with 1 degree of freedom.

The p-value is less than 0.05 (level of significance), so we reject the null hypothesis that there is no association between diabetes and the occurrence of acute myocardial infarction.

Measures of Association

Measures of association are used to assess the strength of an association. For the 2×2 table two measures of association are typically calculated, such as i) relative risk or risk ratio, and ii) odds ratio.

For example, in our NHANES data we may be interested to compare male and female participants' diabetic condition (yes / no). With two groups (male/female), we have a 2×2 contingency table. We can use the risk ratio or odds ratio to assess the strength of association between gender and the prevalence of diabetes.

Let p_1 denotes the proportion (probability) that a male participant has diabetes while p_2 denotes the proportion (probability) that a female participant has diabetes.

The `relative risk` for comparing proportion successes (diabetes, in this case) in two groups is defined as the ratio probabilities:

$$\text{relative risk} = \frac{p_1}{p_2}.$$

A `relative risk` of 1 indicates that the probability of diabetes does not depend on gender.

The `odds ratio` for comparing the odds of successes in two groups is defined as

$$\text{odds ratio, OR} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

The $OR = 1$ indicates that the odds of diabetes does not depend on gender. When $1 < OR < \infty$, male participants are more likely to have diabetes than female participants. For example, $OR = 2$ indicates that the odds for males to have diabetes is twice the odds for females to have diabetes. When $0 < OR < 1$, male participants are less likely to have diabetes than female participants.

```
library(NHANES)

ctab <- table(NHANES$Gender, NHANES$Diabetes)
ctab
```

```
##
##            No  Yes
##   female 4592  357
##   male   4506  403
```

```
library(epitools)

## Relative Risk
rr<- riskratio.wald(ctab)

rr$measure
```

```
##         risk ratio with 95% C.I.
##          estimate     lower   upper
##   female  1.00000        NA      NA
##   male    1.13805 0.9924731 1.30498
```

```r
## Odds ratio

or <- oddsratio.wald(ctab)

or$measure
```

```
##         odds ratio with 95% C.I.
##          estimate     lower    upper
##   female 1.000000        NA       NA
##   male   1.150396 0.9918778 1.334249
```

The risk (probability) of having diabetes in male is 1.14 times higher than that in females. However, the relative risk is not statistically significant as the null value 1 is included in the confidence interval

The odds of having diabetes in male is 1.15 times higher than that in females. However, the odds ratio is not statistically significant as the null value 1 is included in the confidence interval of the odds ratio

The `odds ratio` is a useful measure of association regardless of how the data are collected. However, it is a useful measure for retrospective or case-control study designs.

The `risk ratio` or `relative risk` is an appropriate measure of association when the data are collected prospectively.