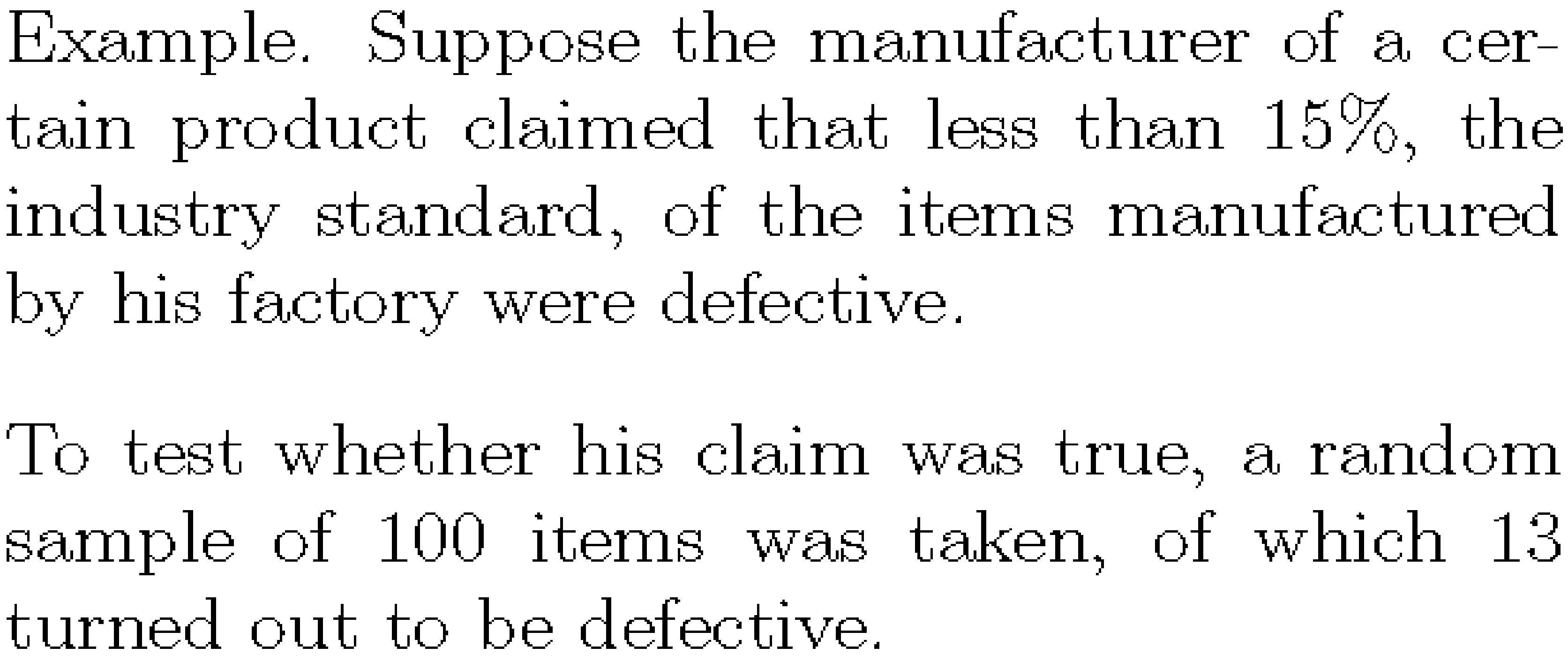
Categorical Data Analysis

Announcement:

Starting next Friday (9/26/2014)
STAT W4201 will meet in
309 Havemeyer



Example. Suppose the manufacturer of a certain product claimed that less than 15%, the industry standard, of the items manufactured by his factory were defective.

To test whether his claim was true, a random sample of 100 items was taken, of which 13 turned out to be defective.

Test the relevant hypothesis.

Let p denote the probability of success

The hypothesis of interest is

$$H_0 : \; p = 0.15$$

vs

$$H_1 : \; p < 0.15$$

A reasonable test is one which rejects H_o for small values of X.

An *exact test* is obtained computing the p-value as $Pr[X \leq 13 \mid p = 0.15]$, or

$$= \sum_{j=1}^{13} \binom{100}{j} 0.15^j 0.85^{n-j}$$

```
binom.test(13, 100, 0.15, alt = "l")
        Exact binomial test

data:  13 out of 100
number of successes = 13, n = 100,
p-value = 0.3474
```

Let X be the number of successes in n trials.

For large n, such that $min(np, nq) \geq 5$,

$$Z = \frac{x - np - 0.5}{\sqrt{npq}}$$

approximately Standard Normal.

```
> prop.test(13,100,0.15,alt="l")
1-sample prop test with continuity correction

X-square = 0.1765, df = 1, p-value = 0.3372

95 percent confidence interval:
0.0000000 0.2009056
```

Denote $\hat{p} = X/n$.

An approximate $100(1-\alpha)\%$ confidence interval for p may be constructed based on the pivot

$$\frac{|\, \hat{p} - p \,| - \frac{1}{2n}}{\sqrt{pq/n}} \leq Z_{\alpha/2}$$

which gives (P_L, P_U), where

$$P_L = \frac{(2n\hat{p} + Z^2_{\alpha/2} - 1) - Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} - (2 + 1/n) + 4\hat{p}(n\hat{q} + 1)}}{2(n + Z^2_{\alpha/2})}$$

and

$$P_U = \frac{(2n\hat{p} + Z^2_{\alpha/2} + 1) + Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + (2 + 1/n) + 4\hat{p}(n\hat{q} - 1)}}{2(n + Z^2_{\alpha/2})}$$

Exact values for P_L and P_U may be obtained by solving the equations

$$Pr[X \geq x \mid P_L] = \alpha/2$$

and

$$Pr[X \leq x \mid P_U] = \alpha/2$$

Example. Suppose the manufacturer of a certain product claimed that the percentage of defective items manufactured by his factory was less than that for the competitor.

To test whether his claim was true, random samples of 131 and 281 items were taken (from each company) of which 161 and 271, respectively, turned out to be non-defective.

Test the relevant hypothesis.

Let p_1 and p_2 denote the respective proportions of defective items.

The hypotheses of interest are

$$H_o: \; p_1 = p_2$$

vs.

$$H_1 : p_1 \neq p_2$$

A large sample test, with Yates' correction for continuity, is given by

$$Z = \frac{\mid \hat{p}_1 - \hat{p}_2 \mid - \frac{1}{2}(\frac{1}{n} + \frac{1}{m})}{\sqrt{\hat{p}_c \hat{q}_c (\frac{1}{n} + \frac{1}{m})}}$$

$$\hat{p}_c = \frac{n\hat{p}_1 + m\hat{p}_2}{n+m}$$

An approximate $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm \left(Z_{\alpha/2}\sqrt{\hat{p}_1\hat{q}_1/n + \hat{p}_2\hat{q}_2/m} + \frac{1}{2}\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

Example (cont’d):

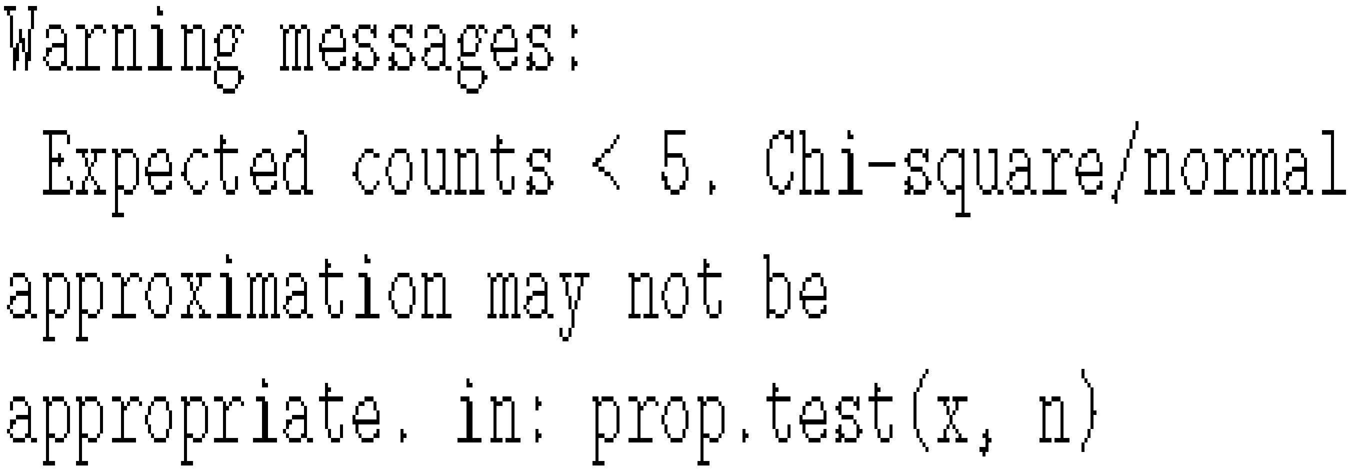
2-sample test for equality of proportions with continuity correction

```
> x <- c(161,131)
> n <- c(271,281)

> prop.test(x,n)

X-square = 8.5518, df = 1, p-value = 0.0035

alternative hypothesis: two.sided

95 percent confidence interval:
0.04169418 0.21411336
```

Example: Suppose random samples of sizes n=100 and m=150 gave x=5 and y=7, respectively.

```
X<-c(5,7); n <- c(100,150)
prop.test(x, n)
```

```
X-square = 0.0146, df = 1, p-value = 0.9039
```



```
Warning messages:
 Expected counts < 5. Chi-square/normal
approximation may not be
appropriate. in: prop.test(x, n)
```

Fisher's Exact Test

Suppose an experiment on the effect of a certain chemical on the mood of subjects (e.g., Depressed/Not Depressed) gave the following data in 3 males and 4 females:

	Depressed	Not Depressed
Male	1	2
Female	3	1

Consider the 2 × 2 table

	Depressed	Not Depressed	
Male	n_{11}	n_{12}	n_{1+}
Female	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

Fix the marginal totals and compute the probably of observing the given cell frequencies:

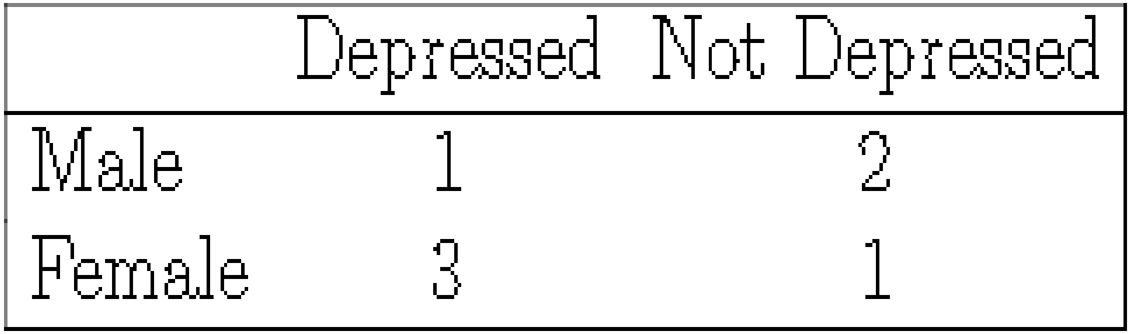
$$p_o = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}}$$

Next compute p^* the probabilities for all tables having the same marginal totals. The p-value is computed as the

$$p = \sum_{p^*:p^*\leq p_o} p*$$

Remarks

- Inference may not be extended to all experiments giving different marginals (i.e., fixing only n)
- Extensions to higher layouts available.



	Depressed	Not Depressed
Male	1	2
Female	3	1

```
depress.data <- matrix(c(1, 2, 3, 1), 2, 2)
fisher.test(depress.data)
p-value = 0.4857
```

Example. Suppose in a survey of public opinion about a certain political issue, a random sample of registered voters taken, n fixed. Sample included voters from each political group: Democrat, Republican, and Other, giving the data displayed below.

	Favor	Do Not Favor	Total
Democrat	198	202	$n_{1+} = 400$
Republican	140	210	$n_{2+} = 350$
Other	133	217	$n_{3+} = 350$
Totals	$n_{+1} = 471$	$n_{+2} = 629$	n = 1100

Hypothesis of interest:

H_o: No association between party Affiliation and Opinion

vs

H_1: There is association

Let p_{ij} denote the probability corresponding to the ij'th cell.
Then under H_0, $p_{ij} = p_{i+}p_{+j}$, and is estimated by

$$\hat{p}_{ij} = \frac{n_{i+}}{n}\frac{n_{+j}}{n}$$

The corresponding expected number is given by

$$E_{ij} = \frac{n_{i+}n_{+j}}{n}$$

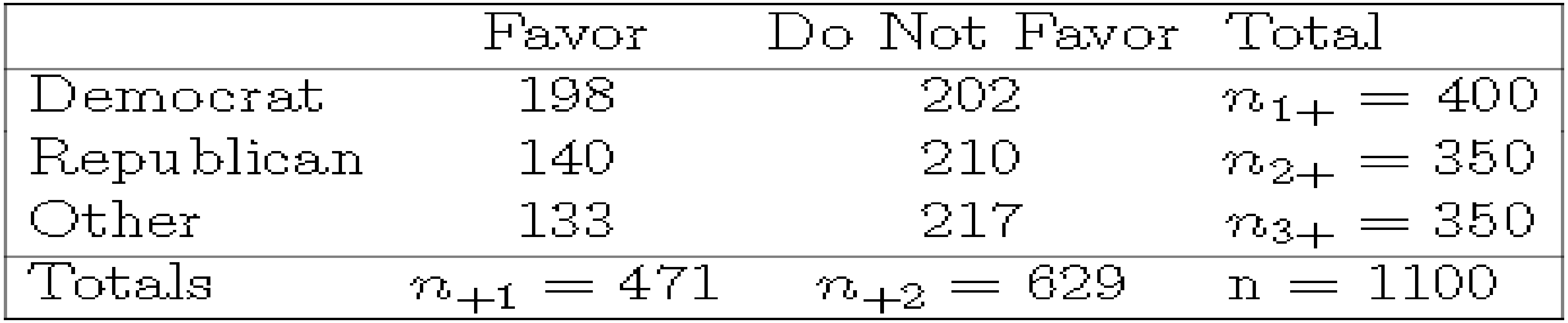
A reasonable test is given by

$$X^2 = \sum \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

and has an approximate $\chi^2_{(I-1)(J-1)}$distribution

The test is commonly referred to as Pearson's chi-square test.

For 2×2 tables, Yates' correction for continuity may be applied.



	Favor	Do Not Favor	Total
Democrat	198	202	$n_{1+} = 400$
Republican	140	210	$n_{2+} = 350$
Other	133	217	$n_{3+} = 350$
Totals	$n_{+1} = 471$	$n_{+2} = 629$	n = 1100

```
opinion.data <- matrix(c(198, 140, 133, 202,
210, 217), 3, 2, byrow = F)

chisq.test(opinion.data)
Pearson's chi-square test without Yates'
continuity correction

X-square = 11.7478, df = 2, p-value = 0.0028
```

Remarks

- Large sample approximation may not be good if the expected cell counts are too small (*ie*, $<$ 5).

- Interpretation of the results is dependent on the sampling scheme (i.e., whether column or row totals are held fixed).

- An alternative to Pearson's chi-square test is the likelihood ratio test. Let
$$\lambda = \frac{ML\ under\ H_0}{ML\ when\ p_{ij}\ are\ unstricted}$$
Put
$$G^2 = -2ln(\lambda)$$
G^2 has, under the null, an approximate $\chi^2_{(I-1)(J-1)}$ distribution.

- Both X^2 and G^2 only tell presence of association, but not the strength of association.

- Both tests depend on the row and column marginal totals, n_{i+}, n_{+j}, and not on the ordering of the rows (or columns).
  Information may be lost if at least one of the variables is ordinal.

- When I or J is large, approximation is generally better for X^2 compared to G^2, even if some $n_{ij} = 1$.

Example. Consider the following artificial data on the relationship between lung cancer and passive smoking.

	Passive	Not Passive
Cancer	281	235
No Cancer	210	279

Application of the Pearson chi-square test gives a p-value = 0.0003.

The effect of passive smoking on cancer is *confounded* with the smoking status of the individual.

	Smoker		Non-smoker	
	Passive	Not Passive	Passive	Not Passive
Cancer	261	118	20	117
No Cancer	130	124	80	155

Application of a chi-square test to each table is not a viable option.

First, it does not draw strength from the combined data, and hence may be less sensitive. The overall level of significance may be inflated, if each test is performed at the usual $\alpha = 0.05$.

Given K 2 × 2 tables, let n_{ijk} be the number of events in the ij'th cell of the k'th table.

The Mantel-Haenszel test is given by

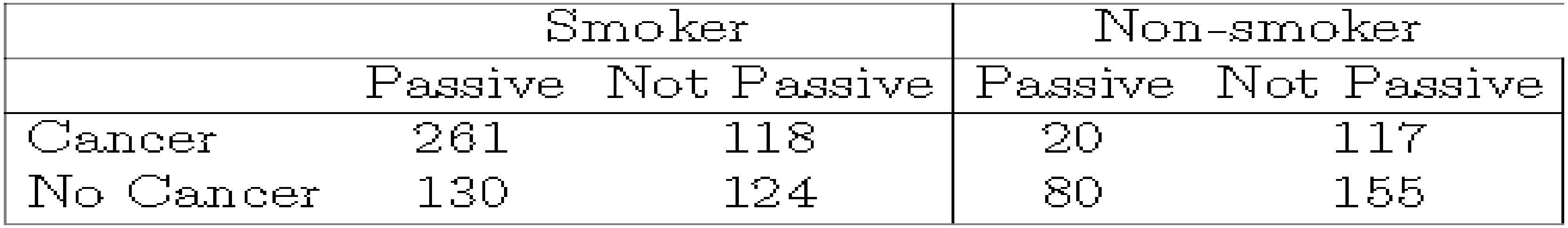
$$X^2_{MH} = \frac{[\mid \Sigma_k(n_{11k} - E_{11k}) \mid -c]^2}{V_{11k}}$$

$$E_{11k} = E[N_{11k} \mid H_0],$$

$$E_{11k} = \frac{n_{1+k}n_{+1k}}{n_{++k}}$$

$$V_{11k} = \sum_k \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n^2_{++k}(n_{++k} - 1)}$$

For large sample, X^2_{MH} has an approximate χ^2_1 distribution.

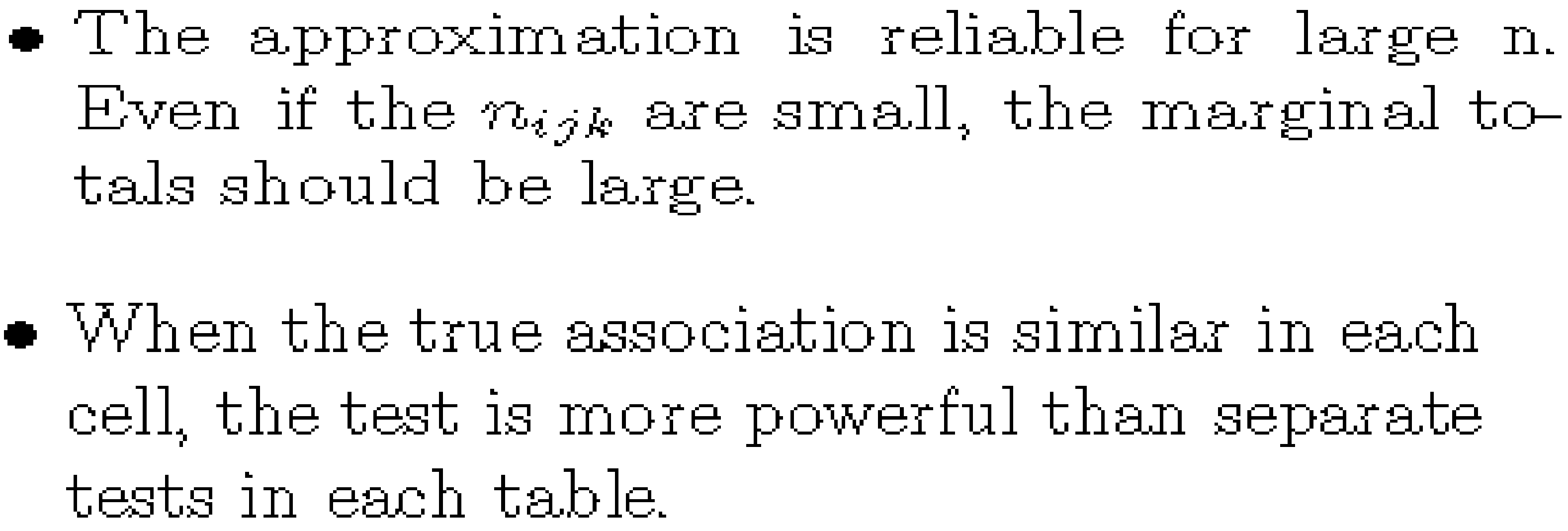


	Smoker		Non-smoker	
	Passive	Not Passive	Passive	Not Passive
Cancer	261	118	20	117
No Cancer	130	124	80	155

```
passive.smoker.data <- array(c(261, 130, 118, 124,
20, 80, 117, 155), c(2, 2, 2))

mantelhaen.test(passive.smoker.data)
```

```
Mantel-Haenszel chi-square = 1.7258, df = 1,
p-value = 0.189
```

Remarks



- The approximation is reliable for large n. Even if the n_{ijk} are small, the marginal totals should be large.
- When the true association is similar in each cell, the test is more powerful than separate tests in each table.

- Interpretation of results when results are not consistent across tables

  NB: p-value still valid, even when tables not homogenous.

Tests for homogeneity across tables.

- The Breslow-Day test, performs reliably when the sample size is large.
- For small samples, an alternative test may be Zelen's procedure.

Extensions to the case of $I \times J \times K$ tables, and when the rows and/or columns are ordinal.

1. *When Both Rows and Columns are Nominal*

   The generalized Cochran-Mantel-Haenszel test for general association concerns the hypothesis

   $$H_o : No\ association\ between\ X\ and\ Y$$

   and the test statistic has an approximate $\chi^2_{(I-1)(J-1)}$ distribution.

2. *When the Row Variable is Nominal and Y Ordinal*

The hypothesis of interest is

Ho : No Diference among row mean scores

and the test statistic has an approximate χ^2_{I-1} distribution.

3. *When Both Row and Column Variables are Ordinal*

The hypothesis of zero correlation is based on

$$M^2 = (n-1)r^2$$

where r is the correlation coefficient between the scores of the row and column. The test statistic has an approximate χ_1^2 distribution.

4. *When X is Ordinal and the Column Variable is Nominal*

   When J=2 and K=1, this reduces to Cochran-Armitage test for trend.

The SAS procedure PROC FREQ implements most of the above situations.

Example. Suppose two eye treatments, A and B, are to be compared with respect to a binary outcome (cure/failure). One hundred eligible subjects, and one eye from each pair randomly assigned to either A or B. The results are given below:

	Cured	Not Cured
A	48	52
B	30	70

Let p_A and p_B be the proportions of cures for treatments A and B, respectively. The null hypothesis of interest is:

$$H_o : p_A = p_B$$

```
> prop.test(c(48,30),c(100,100))

        2-sample test for equality of proportions with
continuity correction

data:  c(48, 30) out of c(100, 100)
X-squared = 6.074, df = 1, p-value = 0.01372
alternative hypothesis: two.sided
95 percent confidence interval:
 0.03712658 0.32287342
sample estimates:
prop 1 prop 2
  0.48   0.30
```

Matched Samples

Due to dependence within each pair, the usual Pearson chi-square test is not applicable to the above table.

Instead, we need to consider the following table for the matched pairs:

		B	
		Cured	Not Cured
A	Cured	8	40
	Not Cured	22	30

To fix ideas, consider the following table:

		B	
		Cured	Not Cured
A	Cured	a	b
	Not Cured	c	d

The corresponding estimators for p_A and p_B are

$$\hat{p}_A = \frac{a+b}{n}$$

and

$$\hat{p}_B = \frac{a+c}{n}$$

Then

$$\hat{p}_A - \hat{p}_B = \frac{b-c}{n}$$

Under H_o, $b \approx c$, so that
$b \sim$ binomial (b+c, $\frac{1}{2}$).

Hence, a test statistic with continuity correction is

$$X^2_{McN} = \frac{[|\, b - c\, | - 1]^2}{b + c}$$

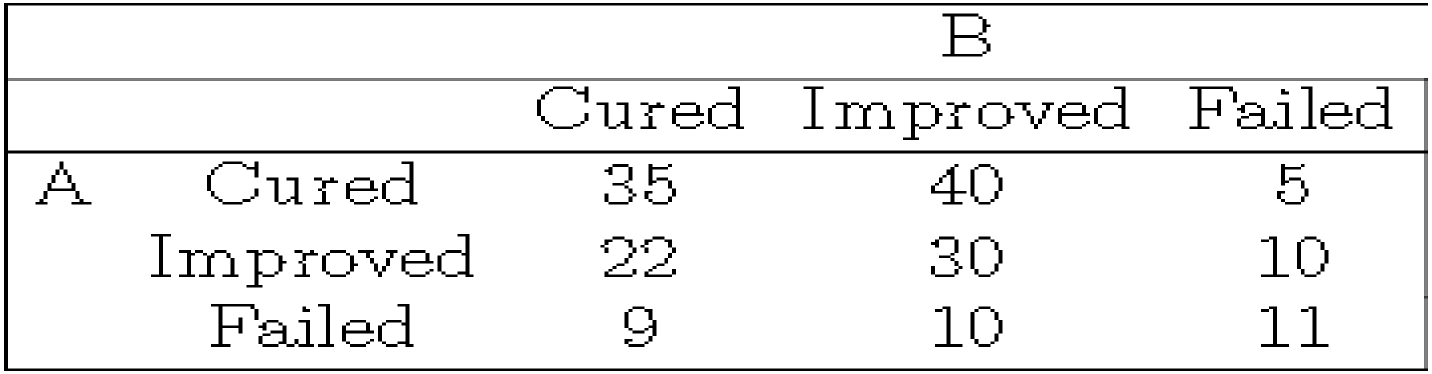
which has an approximate χ^2_1 distribution under H_o.

The test is known as McNemar's test.

		B	
		Cured	Not Cured
A	Cured	8	40
	Not Cured	22	30

```
paired.data <- cbind(c(8, 22), c(40, 30))
mcnemar.test(paired.data)

McNemar's chi-square test with
continuity correction

data:  paired.data
McNemar's chi-square = 4.6613, df = 1,
 p-value = 0.0308
```

Matched Pairs with More Than Two Outcomes



		B		
		Cured	Improved	Failed
A	Cured	35	40	5
	Improved	22	30	10
	Failed	9	10	11

Let n_{ij} denote the number of pairs falling in the ij'th cell, $i, j = 1, 2, \cdots, R$.

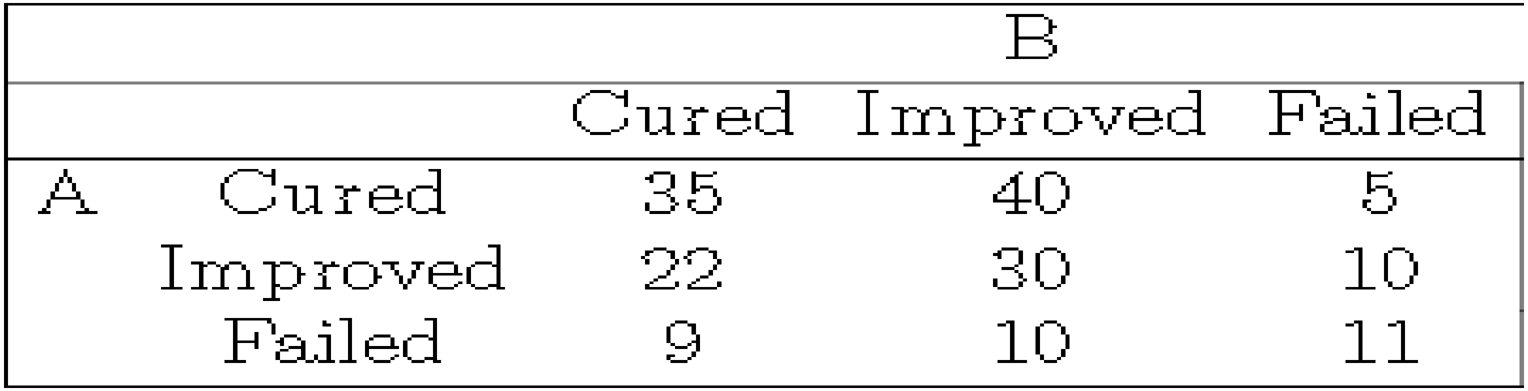
Let p_{ij} be the corresponding proportion. Then the null hypothesis of interest is:

$$H_o : p_{ij} = p_{ji}$$

Test Statistics:

$$X^2_{MC} = \sum_{i:j>i} \frac{(| n_{ij} - n_{ji} | - c)^2}{n_{ij} + n_{ji}}$$

which has a $\chi^2_{R(R-1)/2}$ approximate distribution under Ho..



		B		
		Cured	Improved	Failed
A	Cured	35	40	5
	Improved	22	30	10
	Failed	9	10	11

```
paired.data2 <- cbind(c(35,22,9),c(40,30,10),
c(5,10,11))

> mcnemar.test(paired.data2)

McNemar's chi-square = 6.3687, df = 3,
p-value = 0.095
```

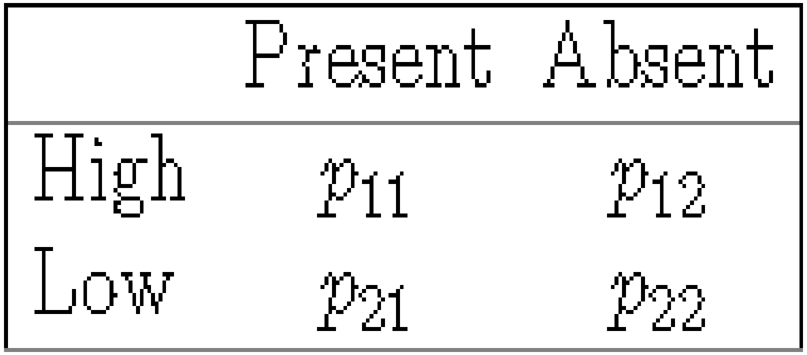
Types of Studies

Interested in evaluating the relationship between:

- Social status (X=0, if low; 1 if high) and
- Mental disease (Y=0, if absent and 1 if present).

Y

X



	Present	Absent
High	p_{11}	p_{12}
Low	p_{21}	p_{22}

Types of Studies

- Cross-Sectional Study
  - Draw a random sample of n individuals,
  - Classify individuals according to their X and Y values.
  - Drawback: Individuals may be systematically excluded from the study, thereby introducing bias.

- Prospective Study
  - Randomize subjects to X=1 and X=0
  - Follow up for a specified period of time.
  - Observe Y at the end of the study
  - Limitation: Cost and time.

Types of Studies (cont’d)

- Retrospective Study
  - Determine samples from Y=1, and a control sample from Y=0 (matched according to certain criteria).
  - Then "look back" to see how many in each category have X=1 and X=0

- Limitation: Since subjects are selected according to the outcome values, and not X, the sample may not be representative of the study population.

Measuring Degree of Association

Consider the following two hypothetical cases.

In one study $p_{11} = 0.01$ and $p_{21} = 0.001$, giving a difference $\triangle = 0.009$.

In the second case, assume $p_{11} = 0.410$ and $p_{21} = 0.401$, giving the same difference $\triangle = 0.009$

However, the relative value $\frac{p_{11}}{p_{21}}$ for the first case is 10, while it is approximately 1 in the second case.

Relative Risk

$$RR = \frac{\hat{p}_{11}}{\hat{p}_{21}}$$

95% Confidence Interval for true RR:

$$ln(\frac{\hat{p}_{11}}{\hat{p}_{21}}) \pm Z_{\alpha/2}\sqrt{\frac{1-\hat{p}_{11}}{n_{1+}\hat{p}_{11}} + \frac{(1-\hat{p}_{21}}{n_{2+}\hat{p}_{21}}}$$

Odds Ratio

Recall that $\frac{p_{11}}{1-p_{11}}$ and $\frac{p_{21}}{1-p_{21}}$ correspond to the odds of having the disease, for X=1 and X=0, respectively. Hence the odds ratio of having the disease for X=1 relative to X=0 is given by

$$\psi = \frac{p_{11}}{1-p_{11}} \div \frac{p_{21}}{1-p_{21}}$$

An estimator of ψ is

$$\hat{\psi} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

A large sample confidence interval for $ln(\psi)$ is

$$ln(\hat{\psi}) \pm Z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

$$\psi = RR\frac{1-p_{21}}{1-p_{11}}$$

If p_{11} and p_{21} are ≈ 0, then ψ may be used to approximate RR.

Remarks:
In case control studies, number of cases (Y=1) and that of controls (Y=0) are controlled by design.

So, cannot compute probabilities corresponding to Y e.g., Pr[Y = 1 / X = 1].

However, we can compute probabilities, like $Pr[X = 1 \mid Y = 1]$.
Since ψ is not affected by interchanging rows and columns, we can estimate ψ and use it to approximate RR, provided p_{11} and p_{21} are small

For three-way tables, the commom odds ratio is estimated as a weighted avearge of the table specific odds ratios.

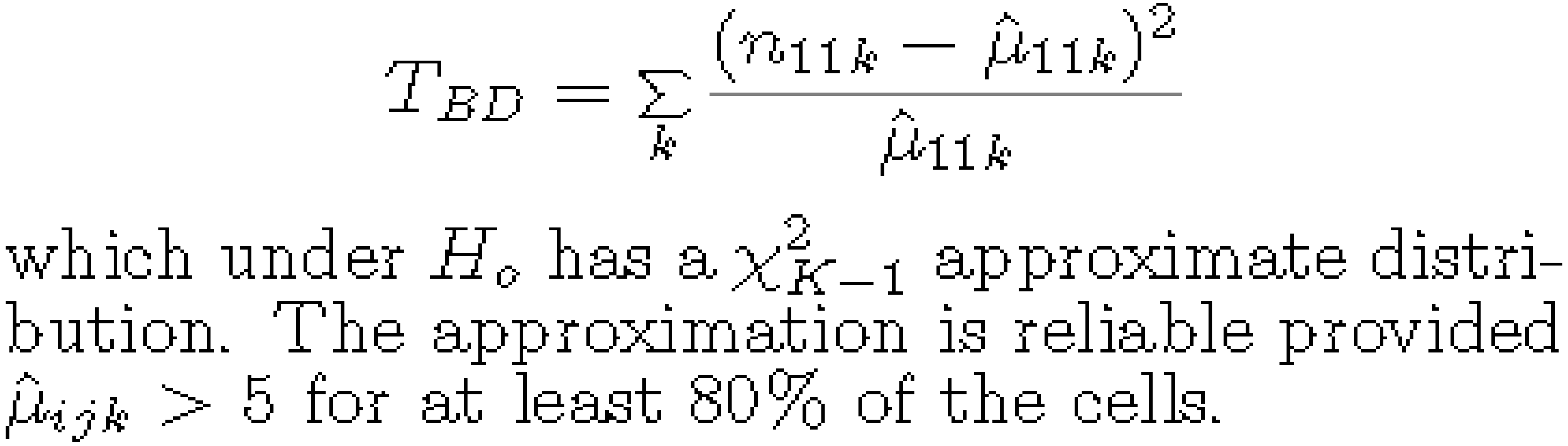
$$\hat{\psi}_{MH} = \sum_k w_k \hat{\psi}_k$$

where

$$w_k = \frac{n_{12k} n_{21k} / n_{++}}{\Sigma_k \, n_{12k} / n_{++k}}$$

and $\hat{\psi}_k$ is the odds ratio for the k'th table.

The Breslow-Day test for homogeneity of the odds ratios:



$$T_{BD} = \sum_k \frac{(n_{11k} - \hat{\mu}_{11k})^2}{\hat{\mu}_{11k}}$$

which under H_o has a χ^2_{K-1} approximate distribution. The approximation is reliable provided $\hat{\mu}_{ijk} > 5$ for at least 80% of the cells.

using the Woolf test for interaction:

```
woolf <- function(x) {
  x <- x + 1 / 2
  k <- dim(x)[3]
  or <- apply(x, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
  w <-  apply(x, 3, function(x) 1 / sum(1 / x))
  1 - pchisq(sum(w * (log(or) - weighted.mean(log(or), w)) ^ 2), k - 1)
}
woolf(UCBAdmissions)
## => p = 0.003, indicating that there is significant heterogeneity.
## (And hence the Mantel-Haenszel test cannot be used.)
```

Problem Set 4

Reading Assignment: Chapters 18 and 19, Ramsey and Schafer

Consider the *survey* data in the R package MASS, consisting of responses of 237 Statistics I students at the University of Adelaide to a number of questions.

1) Suppose we are interested in studying the relationship between gender and smoking status.
   a) Using a suitable test criterion, determine whether there is association between Sex and Smoking status.
   b) Discuss the design of this study, and any potential limitations.
2) Suppose PULSE is defined as HIGH, if the value of Pulse is > 80; LOW, if the value is < 65; and MEDIUM, otherwise.
   a) Determine whether there is association between PULSE and Smoking status, controlling for Sex and Exer.
   b) Assess whether it is appropriate to pool across Sex.
   c) Discuss the impacts of missing values, and any remedial measures. Use Pulse[!is.na(Pulse)] to get non-missing values)

3) Consider now the sub-group of students with Exer value of “None”.
   a) Determine whether there is association between PULSE and Smoking status in this sub-group. For this exercise, classify PULSE as binary (i.e., LOW = Pulse < 80, HIGH, Otherwise), and Smoker as NEVER vs. EVER smoked.
   b) Comment on the limitations of your analysis.

	Time 1	Time 2	Time 3	Wait List
3-Oct	Group 36	OPEN	OPEN	
10-Oct	Group 1	Group 16	Group 28	Group 29
17-Oct	Group 8	Group 15	Group 18	Group 22, 23,9
24-Oct	Group 2	Group 4	Group 7	Group 32, 35
31-Oct	Group 5	Group 12	Group 14	Group 25,34
7-Nov	Group 10	Group 17	Group 24	Group 31
14-Nov	Group 3	Group 13	Group 19	Group 20
21-Nov	Group 11	Group 21	Group 26	Group 6
5-Dec	Group 27	Group 38	OPEN	

Not Signed up: Group 30,33,37