Categorical Data

Section I Overview

(1) Types of Variables

(2) Comparing (2) Categorical Variables

- ∘ As Exploratory Data Analysis (EDA)
- ∘ Confirmatory Data Analysis (CDA)
  - ⋆ Test of Homogeneity
  - ⋆ Test of Independence

(3) 2×2 Tables

- ∘ Sampling designs
- ∘ Testing for association
- ∘ Estimation of "effects"
- ∘ Paired binary data
- ∘ Small sample methods:
  - ⋆ continuity corrections
  - ⋆ Fisher's exact test
- ∘ Using STATA

(4) Stratified Tables

- Confounding
- Effect modification
- Testing for a common OR
  - ⋆ Mantel-Haenszel Test
- Estimation of a common OR
  - ⋆ Woolf's method
  - ⋆ Mantel-Haenszel

(5) Testing for a Trend in Proportions

(6) Measures of Accuracy and Agreement

- Sensitivity & Specificity
- ROC curves
- Kappa statistics

Scales of Measurement

- Nominal
  - Order of categories irrelevant, e.g. gender, color
- Ordinal
  - Order of categories crucial, e.g. "better", "same", "worse"
- Interval (quantitative)
  - Arbitrary origin (0 point) and scale
  - Temperature (°F *vs* °C)
- Ratio (quantitative)
  - Fixed origin, arbitrary scale
  - Distance (miles *vs* kilometers), elapsed time

Factors and Contingency Tables

Definition: A **factor** is a categorical (discrete) variable taking a small number of values that represent the **levels** of the factor

Examples

- Gender with two levels: 1 = Male and 2 = Female
- Disease status with three levels: 1 = "Progression", 2 = "Stable", 3 = "Improved"

Factors may be nominal, ordinal or quantitative

- Quantitative factors often arise by **grouping** of continuous variables into categories, e.g., AgeFactor with 4 levels: 1 = 20-29 yrs, 2 = 30-39, 3 = 40-49, 4 = 50-59

Factors and Contingency Tables

Data description is facilitated by one-way, two-way or multi-way **tables** of **frequencies** of factor levels and their combinations

- To assess whether two factors are related, we often construct a $R \times C$ table that **cross-classifies** the observations according to the 2 factors.
- Examining two-way tables of Factor A *vs* Factor B at each level of a third Factor C shows how the A/B association may be **explained** or **modified** by C

Tests: We can test whether the factors are related using a χ^2 test. Depending on the hypotheses / design we may use

1. $\boxed{\chi^2 \text{ Test of Independence}}$
2. $\boxed{\chi^2 \text{ Test of Homogeneity}}$

We will consider the special case, 2×2 tables, in detail.

Categorical Data

Example 1: Education versus willingness to participate in a study of a vaccine to prevent HIV infection if the study was to start tomorrow. Counts, row percents in each response category, and row totals are given.

	definitely not	probably not	probably	definitely	Total
< high school	52	79	342	226	699
	7.4%	11.3%	48.9%	32.3%	
high school	62	153	417	262	894
	6.9%	17.1%	46.6%	29.3%	
some college	53	213	629	375	1270
	4.2%	16.8%	49.5%	29.5%	
college	54	231	571	244	1100
	4.9%	21.0%	51.9%	22.2%	
some post college	18	46	139	74	277
	6.5%	16.6%	50.2%	26.7%	
graduate/prof	25	139	330	116	610
	4.1%	22.8%	54.1%	19.0%	
Total	264	861	2428	1297	4850
	5.4%	17.8%	50.1%	26.7%	
	Chi-square for Independence: 89.7; df = 15; $p < 0.001$				

Categorical Data

Example 2: from the 1984 General Social Survey

	Job Satisfaction				
Income	Very Dissatisfied	Somewhat Dissatisfied	Moderately Satisfied	Very Satisfied	Total
< 6000	20	24	80	82	
6000-15000	22	38	104	125	
15000-25000	13	28	81	113	
> 25000	7	18	54	92	
Total					

Categorical Data

Example 3: from Doll and Hill (1952) - retrospective assessment of smoking frequency. The tables displays the daily average number of cigarettes for lung cancer patients and control patients.

	Daily # of cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	7 0.5%	55 4.1%	489 36.0%	475 35.0%	293 21.6%	38 2.8%	1357
Control	61 4.5%	129 9.5%	570 42.0%	431 31.8%	154 11.3%	12 0.9%	1357
Total	68	184	1059	906	447	50	2714

Categorical versus Continuous Data

Assume instead of using a 2-sample t-test you grouped the continuous data into C categories. The following table shows the grouped data for $C = 5$.

	G_1	G_2	G_3	G_4	G_5	Total
sample 1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	N_1
sample 2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	N_2

For example:

- Low versus high birth weight ($C = 2$)
- Age categories (decades)

Q: How much information is lost by grouping?

The following table compares the efficiency with which the difference between the groups can be estimated relative to $\overline{Y}_1 - \overline{Y}_2$.

# Categories	Relative Efficiency
2	56.5
3	72.7
4	80.0
5	84.1
6	86.7
7	88.4
8	89.7
9	90.7

NOTE: not exactly the χ^2 we'll discuss since the order is exploited.

Test of Homogeneity

Example #3 is a case where we may want to test whether the smoking frequency is the same for each of the populations sampled. Here we test whether the **groups** are **homogeneous** with respect to a characteristic (factor).

$$H_O \; : \; \text{cancer and control smoking frequency is same}$$
$$H_A \; : \; \text{cancer and control smoking frequency is not the same}$$

Q: What does H_O predict that we observe if we knew the marginal totals only?

	Daily # of cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer							1357
Control							1357
Total	68	184	1059	906	447	50	2714

Test of Homogeneity

A: H_O predicts the following **expectations**:

	Daily # of cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	34	92	529.5	453	223.5	25	1357
Control	34	92	529.5	453	223.5	25	1357
Total	68	184	1059	906	447	50	2714

Note: Each group has the same **proportion** in each cell as the overall **marginal proportion**. The "equal" expectated number for each group is a result of the equal sample sizes (1357 and 1357).

Test of Homogeneity

If each of the **observed counts** were a **Poisson** random variable then we could standardize each cell count if we knew the expectation

$$Z = \frac{O_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}}}$$

Or since a squared normal is a $\chi^2(1)$,

$$\frac{(O_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2(1)$$

However, we don't know the μ_{ij}, but rather we **estimate** them based on the margins.

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Summing all of these squared differences between the **observed** and the **expected** counts provides an overall summary as to the adequecy of H_O.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

X^2 **is known as Pearson's Chi-square Statistic**.

In example #3 the contributions to the χ^2 statistic are:

	Daily # of cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	$\frac{(7-34)^2}{34}$	$\frac{(55-92)^2}{92}$	$\frac{(489-529.5)^2}{529.5}$	$\frac{(475-453)^2}{453}$	$\frac{(293-223.5)^2}{223.5}$	$\frac{(38-25)^2}{25}$	
Control	$\frac{(61-34)^2}{34}$	$\frac{(129-92)^2}{92}$	$\frac{(570-529.5)^2}{529.5}$	$\frac{(431-453)^2}{453}$	$\frac{(154-223.5)^2}{223.5}$	$\frac{(12-25)^2}{25}$	
Total							

	Daily # of cigarettes						
	None	< 5	5-14	15-24	25-49	50+	Total
Cancer	21.44	14.88	3.10	1.07	21.61	6.76	
Control	21.44	14.88	3.10	1.07	21.61	6.76	
Total							

The X^2 statistic is the sum of these contributions:

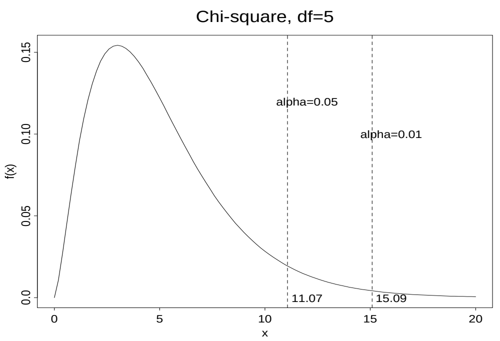
$$X^2 \quad = \quad \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 137.7$$

Test of Homogeneity

Finally, under the null hypothesis of homogeneity, X^2 has approximately a χ^2 distribution. The degrees of freedom is determined by the number of rows (R) and number of columns (C) in the table:

$$df = (R-1) \times (C-1)$$

In example #3 we have $R = 2$, $C = 6$ so under the null hypothesis $X^2 \sim \chi^2(5)$. We use this distribution to determine the $\alpha = 0.05(0.01)$ critical regions:



Therefore, we conclude that the lung cancer patients and control patients are not equivalent with respect to their frequency of cigarette smoking.

Test of Homogeneity

General Procedure

		Factor Levels				Total
		1	2	...	C	
	1	O_{11}	O_{12}	...	O_{1C}	N_1
Groups	2	O_{21}	O_{22}	...	O_{2C}	N_2
	3	O_{31}	O_{32}	...	O_{3C}	N_3
	$\vdots$					
	R	O_{R1}	O_{R2}	...	O_{RC}	N_R
	Total	M_1	M_2	...	M_C	T

1. Compute the expected cell counts under homogeneity assumption:
$$E_{ij} = \frac{N_i M_j}{T}$$
2. Compute the chi-square statistic:
$$X^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
3. Compare X^2 to $\chi^2(df)$ where
$$df = (R-1) \times (C-1)$$
4. Interpret acceptance/rejection or p-value.

NOTE: The use of the χ^2 is valid if all of the $E_{ij} \geq 5$.

Test of Independence

The **Chi-square Test of Independence** is mechanically the same as the test for homogeneity. The only difference is that the $R \times C$ table is formed based on the levels of 2 factors that are cross-classified. Therefore, the null and alternative hypotheses are different:

$$H_O \quad : \quad \text{The two factors are independent}$$

$$H_A \quad : \quad \text{The two factors are not independent}$$

Independence of the 2 factors implies that each row has the same relative frequencies (or each column has the same relative frequencies).

Example # 1 is a situation where individuals are classified according to two factors. The assumption of independence in this example implies that willingness to participate doesn't depend on the level of education.

	definitely not	probably not	probably	definitely	Total
< high school	52 7.4%	79 11.3%	342 48.9%	226 32.3%	699
high school	62 6.9%	153 17.1%	417 46.6%	262 29.3%	894
some college	53 4.2%	213 16.8%	629 49.5%	375 29.5%	1270
college	54 4.9%	231 21.0%	571 51.9%	244 22.2%	1100
some post college	18 6.5%	46 16.6%	139 50.2%	74 26.7%	277
graduate/prof	25 4.1%	139 22.8%	330 54.1%	116 19.0%	610
Total	264 5.4%	861 17.8%	2428 50.1%	1297 26.7%	4850

Q: Based on the observed frequencies how does the independence hypothesis look?

Q: How would the expected cell frequencies be calculated?

Q: How many degrees of freedom would the chi-square have?

Expected Counts under Independence

	definitely not	probably not	probably	definitely	Total
< high school	38.1	124.1	349.9	186.9	699
high school	48.7	158.7	447.6	239.1	894
some college	69.1	225.5	635.8	339.6	1270
college	59.9	195.3	550.7	294.2	1100
some post college	15.1	49.2	138.7	74.1	277
graduate/prof	33.2	108.3	305.4	163.1	610
Total	264 5.4%	861 17.8%	2428 50.1%	1297 26.7%	4850

Summary χ^2 Tests for $R \times C$ Tables

1. Tests of **homogeneity** of a factor across groups or of **independence** of two factors rely on **Pearson's** X^2 **statistic**.
2. X^2 is compared to $\chi^2(\ (R-1) \times (C-1)\ )$.
3. Expected cell counts should all be larger than 5.
4. We have only considered a global test without using possible factor ordering. Ordered factors permit tests for **trend** (See Agresti, 1990).

NEXT:

- 2×2 tables.
- Measures of association: risk difference, risk ratio, odds ratio.
- Retrospective and prospective sampling.
- Paired binary data.

The Ille-et-Vilaine (Grouped) Data

```
COL       VAR          RANGE/VALUES
----------------------------------------
1-2    Age group      1 = 25-34
       (years)        2 = 35-44
                      3 = 45-54
                      4 = 55-64
                      5 = 65-74
                      6 = 75+

3-4    Alcohol        1 =  0-39
       (gms/day)      2 = 40-79
                      3 = 80-119
                      4 = 120+

5-6    Tobacco        1 =  0- 9
       (gms/day)      2 = 10-19
                      3 = 20-29
                      4 = 30+

7-9    Cases          Number of esphogeal
                      cancer cases

10-12  Controls       Number of
                      population controls
----------------------------------------
```

- Read data into STATA

```
. infile age alc tob freq1 freq0 using
  Ille-et-Vilaine.dat
```



```
. summarize
Variable | Obs    Mean     SD.    Min Max
---------+-----------------------------
     age |  88    3.39   1.65     1    6
     alc |  88    2.45   1.12     1    4
     tob |  88    2.41   1.12     1    4
   freq1 |  88    2.27   2.75     0   17
   freq0 |  88    8.81  12.14     0   60
```

- Put data into better format for tabling

```
. reshape long freq, i(age alc tob) j(cc)
(note:  j = 0 1)
. summarize
Variable | Obs  Mean    SD.     Min  Max
---------+------------------------------
     age | 176   3.39   1.65     1     6
     alc | 176   2.45   1.12     1     4
     tob | 176   2.41   1.12     1     4
      cc | 176    .5     .50     0     1
    freq | 176   5.54   9.37     0    60
```

- Label variable values

```
. label define agelab 1 "25-34" 2 "35-44"
  3 "45-54" 4 "55-64" 5 "65-74" 6 "75+"
. label define alclab 1 "0-39" 2 "40-79"
  3 "80-119" 4 "120+"
. label define toblab 1 "0-9" 2 "10-19"
  3 "20-29"  4 "30+"
. label values age agelab
. label values alc alclab
. label values tob toblab
```

- Look at unfortunate age distribution

```
. tabu age cc [fweight=freq]
      |      cc
  age |    0     1 |  Total
------+------------+----------
25-34 |  115     1 |    116
35-44 |  190     9 |    199
45-54 |  167    46 |    213
55-64 |  166    76 |    242
65-74 |  106    55 |    161
  75+ |   31    13 |     44
------+------------+----------
Total |  775   200 |    975
```


- Drop cases to study population frequencies

```
. drop if cc==1
(88 observations deleted)
. tabu age alc [fweight=freq], chi2
      |          alc
  age | 0-39 40-79 80-119 120+ | Total
------+------------------------+-------
25-34 |   61    45     5     4 |   115
35-44 |   88    76    20     6 |   190
45-54 |   77    61    27     2 |   167
55-64 |   77    62    19     8 |   166
65-74 |   60    28    16     2 |   106
  75+ |   23     8     0     0 |    31
------+------------------------+-------
Total |  386   280    87    22 |   775

Pearson chi2(15) =  29.8667   Pr = 0.012
```

- Not easy to discern pattern

- Study average alcohol and tobacco by age

```
. tabu age  [fweight=freq],  summarize(alc)
              |    Summary of alc
       age |        Mean   Std. Dev.   Freq.
-------+------------------------------
 25-34 |   1.5826087   .73726151    115
 35-44 |   1.7052632   .78171815    190
 45-54 |   1.7245509   .77354349    167
 55-64 |    1.746988   .84341835    166
 65-74 |   1.6226415   .81008912    106
   75+ |   1.2580645   .44480272     31
-------+------------------------------
 Total |   1.6709677   .78511407    775


. tabu age  [fweight=freq],  summarize(tob)
              |     Summary of tob
       age |        Mean   Std. Dev.   Freq.
-------+------------------------------
 25-34 |   1.7652174   1.1028705    115
 35-44 |   1.7421053   .99299973    190
 45-54 |   1.7065868    .8938301    167
 55-64 |   1.6746988   .86841456    166
 65-74 |    1.490566   .74635984    106
   75+ |   1.5806452   .92282875     31
-------+------------------------------
 Total |   1.6825806   .93172261    775
```


2 x 2 Tables

Example 1: Pauling (1971)
Patients are randomized to either receive Vitamin C or placebo. Patients are followed-up to ascertain the development of a cold.

	Cold - Y	Cold - N	Total
Placebo	31	109	140
Vitamin C	17	122	139
Total	48	213	279

Q: Is treatment with Vitamin C associated with a reduced probability of getting a cold?

Q: If Vitamin C is associated with reducing colds, then what is the magnitude of the effect?

Example 2: Rothman and Keller (1972)
Patients with (cases) and without (controls) oral cancer were surveyed regarding their smoking frequency (this table collapses over the smoking frequency categories).

	Case	Control	Total
Smoker	85	26	111
non-Smoker	398	421	819
Total	483	447	930

Q: Is oral cancer associated with smoking?

Q: If smoking is associated with oral cancer, then what is the magnitude of the risk?

Example 3: Norusis (1988)
In 1984, a random sample of US adults were cross-classified based on their income and reported job satisfaction:

	Dissatisfied	Satisfied	Total
$< \$15,000$	104	391	495
$\geq \$15,000$	66	340	406
Total	170	731	901

Q: Is salary associated with job satisfaction?

Q: If salary is associated with satisfaction then what is the magnitide of the effect?

Example 4: HIVNET (1995)
Subjects were surveyed regarding their knowledge of vaccine trial concepts both at baseline and at month 3 after an informed consent process. The following table shows the subjects cross-classified according to the two responses.

		Month 3		
		Incorrect	Correct	Total
Baseline	Incorrect	251	178	429
	Correct	68	98	166
	Total	319	276	595

Q: Did the informed consent process improve knowledge?

Q: If informed concent improved knowledge then what is the magnitude of the effect?

Each of these tables can be represented as follows:

	D	$\overline{D}$	Total
E	a	b	$(a+b)=n_1$
$\overline{E}$	c	d	$(c+d)=n_2$
Total	$(a+c)=m_1$	$(b+d)=m_2$	N

The question of association can be addressed with **Pearson's** χ^2. We compute the **expected** cell counts as follows:

Expected:

	D	$\overline{D}$	Total
E	$\frac{n_1 m_1}{N}$	$\frac{n_1 m_2}{N}$	$(a+b)=n_1$
$\overline{E}$	$\frac{n_2 m_1}{N}$	$\frac{n_2 m_2}{N}$	$(c+d)=n_2$
Total	$(a+c)=m_1$	$(b+d)=m_2$	N

2 x 2 Tables

Pearson's chi-square is given by:

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{4} (O_i - E_i)^2 / E_i \\
&= (a - \frac{n_1 m_1}{N})^2 / (\frac{n_1 m_1}{N}) + (b - \frac{n_1 m_2}{N})^2 / (\frac{n_1 m_2}{N}) + \\
&\quad (c - \frac{n_2 m_1}{N})^2 / (\frac{n_2 m_1}{N}) + (d - \frac{n_2 m_2}{N})^2 / (\frac{n_2 m_2}{N}) + \\
&= \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2}
\end{aligned}
$$

Q: How does this χ^2 test compare in Example 1 to simply using the 2 sample binomial test of
$H_O \; : \; P(D \mid E) = P(D \mid \overline{E})$?

Q: How does this χ^2 test compare in Example 2 to simply using the 2 sample binomial test of
$H_O \; : \; P(E \mid D) = P(E \mid \overline{D})$?

2 x 2 Tables

Example 1: Pauling (1971)

	Cold - Y	Cold - N	Total
Placebo	31	109	140
Vitamin C	17	122	139
Total	48	213	279

Test of Independence:

H_O : treatment is independent of disease

H_A : treatment is not independent of disease

$$
\begin{aligned}
X^2 &= \frac{N(ad - bc)^2}{n_1 n_2 m_1 m_2} \\
&= \frac{279(31 \times 122 - 17 \times 109)^2}{140 \times 139 \times 48 \times 213} \\
&= 5.22
\end{aligned}
$$

For the p-value we compute $P(\chi^2(1) > 5.22) = 0.022$. Therefore, we reject the independence of treatment and disease.

	Cold - Y	Cold - N	Total
Placebo	31	109	140
Vitamin C	17	122	139
Total	48	213	279

Test of 2 binomial proportions:

$$p_1 = P(\text{ cold } \mid \text{ Vitamin C })$$
$$p_2 = P(\text{ cold } \mid \text{ placebo })$$

$$\begin{aligned} H_O &: p_1 = p_2 \\ H_A &: p_1 \neq p_2 \end{aligned}$$

$$\begin{aligned} Z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_o(1-\hat{p}_o)(1/n_1 + 1/n_2)}} \\ &= \frac{(17/139 - 31/140)}{\sqrt{\frac{48}{279}\frac{213}{279}(1/139 + 1/140)}} \\ &= -2.284 \end{aligned}$$

For the 2-sided p-value we compute $2 \times P(|Z| > 2.284) = 0.022$. Therefore we reject H_O **with the exact same result as the χ^2 test**. $(Z^2 = X^2)$

2 x 2 Tables: Applications In Epidemiology

Example 1 fixed the number of E and $\overline{E}$ then evaluated the disease status after a fixed period of time. This is a **prospective study**. Given this design we can estimate the **relative risk**:

$$RR = \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

The range of RR is $[0, \infty)$. By taking the logarithm, we have $(-\infty, +\infty)$ as the range for $\log(RR)$ and a better approximation to normality for the estimated $\log(\hat{RR})$:

$$\begin{aligned} \log(\hat{RR}) &= \log\left(\frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})}\right) \\ &= \log\left(\frac{a/n_1}{c/n_2}\right) \\ \log(\hat{RR}) &\sim N(\ \log(p_1/p_2)\ ,\ \frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}\ ) \end{aligned}$$

	Cold - Y	Cold - N	Total
Placebo	31	109	140
Vitamin C	17	122	139
Total	48	213	279

The estimated relative risk is:

$$\hat{RR} = \frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})} = \frac{31/140}{17/139} = 1.81$$

We can obtain a confidence interval for the relative risk by first obtaining a confidence interval for the log-RR:

$$\log(\hat{RR}) \pm Q_Z^{(1-\frac{\alpha}{2})} \times \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$$

For Example 1, a 95% confidence interval for the log

relative risk is given by:

$$\begin{array}{rcl} \log(\hat{RR}) & \pm & 1.96 \times \sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1} + \frac{1-\hat{p}_2}{\hat{p}_2 n_2}} \\ \log(1.81) & \pm & 1.96 \times \sqrt{\frac{109}{(31)(140)} + \frac{122}{(17)(139)}} \\ 0.593 & \pm & 1.96 \times 0.277 \\ 0.593 & \pm & 0.543 \\ (\ 0.050 & , & 1.116\ ) \end{array}$$

To obtain a 95% confidence interval for the **relative risk** we exponentiate the end-points of the interval for the **log - relative risk**. Therefore,

$$\begin{array}{ccc} (\ \exp(0.050) & , & \exp(1.116)\ ) \\ (\ 1.05 & , & 3.05\ ) \end{array}$$

is a 95% confidence interval for the relative risk.

2 x 2 Tables: Applications In Epidemiology

In Example 2 we fixed the number of **cases** and **controls** then ascertained exposure status. Such a design is known as a **case- control study**. Based on this we are able to directly estimate:

$$P(E \mid D) \text{ and } P(E \mid \overline{D})$$

However, we generally are interested in the relative risk which is not estimable from these data alone - we've fixed the number of diseased and diseased free subjects. Instead of the relative risk we can estimate the **exposure odds ratio** which Cornfield (1951) showed equivalent to the **disease odds ratio**:

$$\frac{P(E \mid D)/(1 - P(E \mid D))}{P(E \mid \overline{D})/(1 - P(E \mid \overline{D}))} = \frac{P(D \mid E)/(1 - P(D \mid E))}{P(D \mid \overline{E})/(1 - P(D \mid \overline{E}))}$$

Odds Ratio

"Mr. Yule has invented a series of statistical methods which are in no case based on a reasoned theory, but which possess the dangerous fascination of very easy and ready application, and therefore are at once seized upon as applicable to all sorts of problems by those who are without adequate training in statistical theory, or without the mathematical requisite to weigh cautiously their logical basis."

"If Mr. Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory..."

Karl Pearson and David Heron
Biometrika, (1913)

Odds Ratio

Furthermore, for rare diseases, $P(D \mid E) \approx 0$ so that the **disease odds ratio** approximates the relative risk:

$$\frac{P(D \mid E)/(1 - P(D \mid E))}{P(D \mid \overline{E})/(1 - P(D \mid \overline{E}))} \approx \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

Since with case-control data we are able to efficiently estimate the exposure odds ratio we are then able to equivalently estimate the disease odds ratio which for rare diseases approximates the relative risk.

2 x 2 Tables: Applications In Epidemiology

Like the relative risk, the odds ratio has $[0, \infty)$ as its range. The **log odds ratio** has $(-\infty, +\infty)$ as its range and the normal approximation is better as an approximation to the estimated log odds ratio.

$$\begin{aligned} OR &= \frac{p_1/q_1}{p_2/q_2} \\ \hat{OR} &= \frac{\hat{p}_1/\hat{q}_1}{\hat{p}_2/\hat{q}_2} \\ \hat{OR} &= \frac{ad}{bc} \end{aligned}$$

Confidence intervals are based upon:

$$\log(\hat{OR}) \sim N(\ \log(OR)\ , \frac{1}{n_1 p_1} + \frac{1}{n_1 q_1} + \frac{1}{n_2 p_2} + \frac{1}{n_2 q_2}\ )$$

Therefore, a $(1 - \alpha)$ confidence interval for the log odds ratio is given by:

$$\log(\frac{ad}{bc}) \pm Q_Z^{(1-\frac{\alpha}{2})} \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Example 2:

	Case	Control	Total
Smoker	85	26	111
non-Smoker	398	421	819
Total	483	447	930

The estimated odds ratio (odds of cancer for smokers relative to the odds of cancer for non-smokers) is given by:

$$\hat{OR} = \frac{85 \times 421}{398 \times 26} = 3.458$$

A 95% confidence interval for the log odds ratio is given by:

$$\log(3.458) \quad \pm \quad 1.96 \times \sqrt{\frac{1}{85} + \frac{1}{26} + \frac{1}{398} + \frac{1}{421}}$$

$$1.241 \quad \pm \quad 1.96 \times 0.235$$

$$1.241 \quad \pm \quad 0.460$$

$$(\ 0.781 \quad , \quad 1.701)$$

To obtain a 95% confidence interval for the odds ratio we simply exponentiate the end-points of the interval for the log odds ratio. Therefore,

$$(\ \exp(0.781) \quad , \quad \exp(1.701)\ )$$

$$(\ 2.184 \quad , \quad 5.479\ )$$

is a **95% confidence interval for the odds ratio**.

2 x 2 Tables: Applications In Epidemiology

Example 3 is an example of a **cross-sectional** study since only the total for the table is fixed in advance. The row totals or column totals are not fixed in advance.

In epidemiological studies, the relative risk or odds ratio may be used to summarize the association when using a X-sectional design. The major distinction from a **prospective** study is that a **cross-sectional** study will reveal the number of cases currently in the sample. This is known as **prevalent** cases. In a prospective study we count the number of new cases, or **incident** cases.

Study	Probability	Description
Cohort	incidence	probability of obtaining the disease
Cross-sectional	prevalence	probability of having the disease

SUMMARY: Sampling Models for 2×2 Tables

Design (cohort study):

- Sample n_1 "exposed" and n_2 "non-exposed" from population
- Follow all subjects an **equal period of time** for occurrence of a health event (response)

Response: Observe r_1 "cases" or "responses" among exposed and r_2 among non-exposed

Sampling Model (two independent binomials):

$$\begin{aligned} r_1 &\sim \text{binom}(p_1, n_1) \\ r_2 &\sim \text{binom}(p_2, n_2) \end{aligned}$$

Parameters:

- $p_1 = \Pr[D|E]$ = disease probability for exposed
- $p_2 = \Pr[D|\bar{E}]$ = disease probability for non-exposed

Examples:

- r_i = numbers of low birthweight (LBW) babies among n_i live births
- r_i = numbers of deaths from coronary heart disease (CHD) in 30 years of follow-up of Framingham subcohorts of size n_i

SUMMARY: Measures of Association for 2 × 2 Tables

RD $= p_1 - p_2 =$ risk difference

- Also known as **attributable** or **excess** risk
- Measures **absolute effect**: number of cases among the
  exposed that are attributable to the exposure

RR $= p_1/p_2 =$ risk ratio

- Measures **relative** effect of exposure
- Constrained by denominator probability:
  - RR ≤ 2 if p_2=0.5
  - RR ≤ 1.25 if $p_2 = 0.8$

SUMMARY: Measures of Association for 2 × 2 Tables

OR $= [p_1(1-p_2)]/[(1-p_1)p_2] =$ odds ratio

- Not constrained by denominator:
  - $0 <$ OR $< \infty$ regardless of p_2
- "Natural" parameter for statistical inference
  - Arises in logistic regression
  - "Asymptotic" p-values and CI's valid even for small
    to moderate sized samples
- Invariant to interchange of rows and cols (changing only one switches OR to 1/OR)
- Approximates RR for rare events
  $(1 - p_1 \approx 1 - p_2 \approx 1)$

Null hypothesis expressed:

$$p_1 = p_2 \Longleftrightarrow \text{RD} = 0 \Longleftrightarrow \text{RR} = \text{OR} = 1$$

SUMMARY: Statistical Inference on RD, RR and OR

Estimation: $\hat{p}_1 = 1 - \hat{q}_1 = r_1/n_1; \quad \hat{p}_2 = 1 - \hat{q}_2 = r_2/n_2$

$$\widehat{\mathrm{RD}} = \hat{p}_1 - \hat{p}_2; \qquad \widehat{\mathrm{RR}} = \frac{\hat{p}_1}{\hat{p}_2}; \qquad \widehat{\mathrm{OR}} = \frac{\hat{p}_1 \hat{q}_2}{\hat{q}_1 \hat{p}_2}$$

Standard errors:

$$\begin{aligned}
\mathrm{SE}(\widehat{\mathrm{RD}}) &= \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{\frac{ab}{(a+b)^3} + \frac{cd}{(c+d)^3}} \\
\mathrm{SE}(\log \widehat{\mathrm{RR}}) &= \sqrt{\frac{\hat{q}_1}{\hat{p}_1 n_1} + \frac{\hat{q}_2}{\hat{p}_2 n_2}} = \sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}} \\
\mathrm{SE}(\log \widehat{\mathrm{OR}}) &= \sqrt{\frac{1}{n_1\hat{p}_1} + \frac{1}{n_1\hat{q}_1} + \frac{1}{n_2\hat{p}_2} + \frac{1}{n_2\hat{q}_2}} \\
&= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}
\end{aligned}$$

Interval estimation: $100(1-\alpha)\%$ CI for parameter θ:

$$(\theta_{\mathrm{L}}, \theta_{\mathrm{U}}) = \left(\hat{\theta} - Z_{1-\alpha/2}\mathrm{SE}(\hat{\theta}), \hat{\theta} + Z_{1-\alpha/2}\mathrm{SE}(\hat{\theta})\right)$$

Back transformation: If θ= log(RR), then

$$(\mathrm{RR}_{\mathrm{L}}, \mathrm{RR}_{\mathrm{U}}) = \left(e^{\theta}\mathrm{L}, e^{\theta}\mathrm{U}\right)$$

SUMMARY: Odds Ratios and Case-Control Studies

In a **case-control** study the marginal totals m_1 (cases) and m_2 (controls) are **fixed by design**. The numbers of exposed cases (a) and exposed controls (b) are **random**, as is their sum $a + b = n_1$:

$$\begin{aligned}
a &\sim \mathrm{binom}(P_1, m_1) \\
b &\sim \mathrm{binom}(P_2, m_2)
\end{aligned}$$

where P_1 and P_2 are the **exposure** probabilities for cases and controls, respectively. These are typically very different from the disease probabilities p_1 and p_2. However, the **exposure odds ratio** and the **disease odds ratio** agree. By **Bayes theorem**:

$$\begin{aligned}
\frac{P_1(1-P_2)}{(1-P_1)P_2} &= \frac{p_1(1-p_2)}{(1-p_1)p_2}, \quad \textit{i.e.}, \\
\frac{\Pr(E|D)\Pr(\bar{E}|\bar{D})}{\Pr(\bar{E}|D)\Pr(E|\bar{D})} &= \frac{\Pr(D|E)\Pr(\bar{D}|\bar{E})}{\Pr(\bar{D}|E)\Pr(D|\bar{E})}
\end{aligned}$$

SUMMARY: Odds Ratios and Case-Control Studies

Conclusions: 1. The exposure odds ratio calculated from a case-control sample estimates the disease odds ratio in the population

2. Inferences on odds ratios are identical under cohort and case-control sampling. Methods developed for one type of sampling may be applied directly to the other, and vice-versa.

Illustration using STATA

```
. use ille-et-vilaine
. summarize
 Var | Obs   Mean    SD.   Min  Max
-----+----------------------------
 age | 176  3.387   1.645    1    6
 alc | 176  2.455   1.121    1    4
 tob | 176  2.409   1.118    1    4
  cc | 176   .5      .501    0    1
freq | 176  5.540   9.366    0   60
```

- Create tobacco “exposure” factor with two levels

```
. recode tobexp 1/2=0 3/4=1
(176 changes made)
. table tob tobexp
----------+-----------
          |   tobexp
      tob |    0     1
----------+-----------
      0-9 |   48
    10-19 |   48
    20-29 |         40
      30+ |         40
----------+-----------
```

Use STATA's `epitab` Commands

```
. cs cc tobexp [fweight=freq], or

                 |   Exp   Unexp  | Total
-----------------+----------------|---------
           Cases |    64     136  |   200
        Noncases |   150     625  |   775
-----------------+------------------------+--------
           Total |   214     761  |   975
                 |                        |
            Risk |  .299    .179  |  .205
                 |                        |
                   Estimate [95% Conf. Interval]
                 |---------+---------------------
 Risk difference |  .120   .053   .187
      Risk ratio | 1.673  1.296  2.160
 Attr. frac. ex. |  .402   .228   .537
 Attr. frac. pop |  .129
      Odds ratio | 1.961  1.388  2.770 (Cornfield)
                 +----------------------------------
      chi2(1) =     14.84  Pr>chi2 = 0.0001
```

Query: Which of these measures of association make **no** sense for these data, and why not?

STATA Commands for Case-Control Data

```
. tabodds cc tobexp [fweight=freq]
--------+-----------------------------------
tobexp  |cases cont  odds  [95% Conf. Interval]
--------+-----------------------------------
     0  |  136  625  0.218    0.181  0.262
     1  |   64  150  0.427    0.318  0.572
------------+--------------------------------
Test of homogeneity : chi2(1)  =    14.82
  (equal odds)        Pr>chi2  =   0.0001

Score test for trend: chi2(1)  =    14.82
                      Pr>chi2  =   0.0001

. tabodds cc tobexp [fweight=freq], or co
--------+----------------------------------------
        |            odds        Cornfield
tobexp  |cases cont  ratio  [95% Conf. Interval]
--------+----------------------------------------
     0  |  136  625  1.000    0.770   1.299
     1  |   64  150  1.961    1.388   2.770
--------+----------------------------------------
```

- Now consider all four levels of tobacco

```
. tabodds cc tob [fweight=freq]
-------+-----------------------------------------
  tob  |cases  cont  odds   [95% Conf. Interval]
-------+-----------------------------------------
  0-9  |   78   447  0.175    0.137   0.222
10-19  |   58   178  0.326    0.242   0.438
20-29  |   33    99  0.333    0.225   0.494
  30+  |   31    51  0.608    0.389   0.950
-----------+-------------------------------------
Test of homogeneity : chi2(3)  =    29.33
  (equal odds)        Pr>chi2  =   0.0000

Score test for trend: chi2(1)  =    26.93
                      Pr>chi2  =   0.0000
```



```
. tabodds cc tob [fweight=freq], or
-------+------------------------------------------
  tob  |Odds ratio  chi2  P>chi2  [95% CI     ]
-------+------------------------------------------
  0-9  | 1.000
10-19  | 1.867      10.46  0.0012  1.271  2.743
20-29  | 1.910       7.72  0.0055  1.200  3.040
  30+  | 3.483      25.31  0.0000  2.074  5.850
-------+-------------------------------------------
Test of homogeneity : chi2(3)  =    29.33
   (equal odds)       Pr>chi2  =   0.0000

Score test for trend: chi2(1)  =    26.93
                      Pr>chi2  =   0.0000
```

Paired Binary Data

Example 4 measured a binary response pre and post treatment. This is an example of **paired binary data**. One way to display these data is the following:

	Correct	Incorrect	Total
Baseline	166	429	595
Month 3	276	319	595
Total	442	748	1190

Q: Can't we simply use χ^2 Test of Homogeneity to assess whether this is evidence for an increase in knowledge?

A: NO!!! The χ^2 tests assume that the rows are **independent** samples. In this design it is the same 595 people at Baseline and at 3 months.

Paired Binary Data

For paired binary data we display the results as follows:

		Time 2	
		0	1
Time 1	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

This analysis explicitly recognizes the heterogeneity of subects. Thus, those that score (0,0) and (1,1) provide no information about treatment since they may be "weak" or "strong" individuals. These are known as the **concordant pairs**. The information regarding treatment is in the **discordant pairs**, (0,1) and (1,0).

$$\begin{aligned} p_1 &= \text{success probability at Time 1} \\ p_2 &= \text{success probability at Time 2} \end{aligned}$$

$$\begin{aligned} H_O &: p_1 = p_2 \\ H_A &: p_1 \neq p_2 \end{aligned}$$

Paired Binary Data: McNemar's Test

Under the null hypothesis, $H_O: \ p_1 = p_2$, we expect equal numbers to change from 0 to 1 and from 1 to 0 ($E[n_{01}] = E[n_{10}]$). Specifically, under the null:

$$\begin{aligned} M &= n_{01} + n_{10} \\ n_{10} \mid M &\sim b(\ M\ ,\ \frac{1}{2}\ ) \end{aligned}$$

$$Z = \frac{n_{10} - M\frac{1}{2}}{\sqrt{M\frac{1}{2}(1-\frac{1}{2})}}$$

Under H_O, $Z^2 \sim \chi^2(1)$, and forms the basis for **McNemar's Test for Paired Binary Responses**.

The odd ratio comparing the odds of success at Time 2 to Time 1 is estimated by:

$$\hat{OR} = \frac{n_{01}}{n_{10}}$$

Confidence intervals can be obtained as described in Breslow and Day (1981), section 5.2, or in Armitage and Berry (1987), chapter 16.

Paired Binary Data

A common epidemiological design is to **match** cases and controls regarding certain factors (age, smoking...) then ascertain the exposure history for each member of the pair. The results for all pairs can be summarized by:

		Case	
		$E-$	$E+$
Control	$E-$	n_{00}	n_{01}
	$E+$	n_{10}	n_{11}

Given this design we can use McNemar's Test to test the hypotheses

$$\begin{aligned} H_O &: \ P(D \mid E) = P(D \mid \overline{E}) \ \ (OR = 1) \\ H_A &: \ P(D \mid E) \neq P(D \mid \overline{E}) \ \ (OR \neq 1) \end{aligned}$$

Example 4:

		Month 3		
		Incorrect	Correct	Total
Baseline	Incorrect	251	178	429
	Correct	68	98	166
	Total	319	276	595

We can test $H_O: \; p_2 = p_1$ using **McNemar's Test**

$$
\begin{aligned}
Z &= \frac{n_{01} - M\frac{1}{2}}{\sqrt{M\frac{1}{2}\frac{1}{2}}} \\
&= \frac{178 - (178+68)/2}{\sqrt{(178+68)/4}} \\
&= 7.01
\end{aligned}
$$

Comparing 7.01^2 to a $\chi^2(1)$ we find that $p < 0.001$. Therefore we reject the null hypothesis of equal success probabilities for Time 1 and Time 2.

We estimate the odds ratio as $\hat{OR} = 178/68 = 2.62$.

SUMMARY: Paired Binary Data

The appropriate data display was as in the first table, as follows:

		Time 2	
		0	1
Time	0	n_{00}	n_{01}
1	1	n_{10}	n_{11}

This approach explicitly recognizes the heterogeneity of subects. Those that score (0,0) and (1,1) provide no information about treatment since they may be "weak" or "strong" individuals. These are known as the **concordant pairs**. The information regarding treatment is in the **discordant pairs**, (0,1) and (1,0).

$$
\begin{aligned}
p_1 &= \text{success probability at Time 1} \\
p_2 &= \text{success probability at Time 2}
\end{aligned}
$$

$$\mathrm{H}_0\text{: } p_1 = p_2 \quad \mathrm{H}_A\text{: } p_1 \neq p_2$$

Under the alternative hypothesis H_A,

$$\Pr\left((1,0)|(1,0) \text{ or } (0,1)\right) = \frac{p_1 q_2}{p_1 q_2 + p_2 q_1} = \frac{\text{OR}}{\text{OR}+1}$$

where OR is the **odds ratio**. Thus we may test the null hypothesis and make inferences about the odds ratio using the **binomial distribution** for the number of discordant pairs of type (1,0) given the total number $n_{10} + n_{01}$ of discordant pairs:

$$n_{10} \sim \text{binom}\left(\frac{\text{OR}}{\text{OR}+1}\ ,\ n_{10} + n_{01}\right)$$

McNemar's Test for Paired Binary Data

Under the null hypothesis $\text{H}_0 : p_1 = p_2$ or OR=1, we expect equal numbers of subjects to change from 0 to 1 and from 1 to 0 so that $\text{E}(n_{01}) = \text{E}(n_{10})$. Specifically:

$$\begin{aligned}
n_{10}|(n_{10} + n_{01}) &\sim \text{binom}\left(\frac{1}{2}\ ,\ n_{10} + n_{01}\right) \\
Z &= \frac{n_{10} - \frac{1}{2}(n_{10} + n_{01})}{\sqrt{(n_{10} + n_{01})\frac{1}{2}(1 - \frac{1}{2})}} \\
Z^2 &= \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}} \\
&= \frac{(178 - 68)^2}{178 + 68} = 49.2
\end{aligned}$$

for Example 4. Under H_0, $Z^2 \sim \chi^2(1)$, is the statistic used for **McNemar's Test** for paired binary responses (p=.00000 in example) .

The odds ratio comparing the odds of success at Time 2 to the odds of success at Time 1 (note switch) is estimated by:

$$\begin{aligned}\widehat{OR} &= \frac{n_{01}}{n_{10}} \\ &= \frac{178}{68} = 2.62\end{aligned}$$

Confidence intervals can be obtained from the binomial distribution involving OR/(OR+1) as described by Breslow and Day (1980) §5.2 or by Armitage and Berry (1987) §16.

The Exact Test and the Continuity Correction

Suppose that in another set of matched pairs data

$$n_{10} = 8, \; n_{01} = 2 \text{ and } n_{10} + n_{01} = 10$$

so that (with the p-value one-sided)

$$\widehat{\mathrm{OR}} = \frac{8}{2} = 4, \quad Z^2 = \frac{36}{10} = 3.6, \quad p = 0.029.$$

The **exact** p-value for these data may be obtained from the upper tail of the **binomial**(1/2,10) distribution as follows:

$$\frac{\binom{10}{8} + \binom{10}{9} + \binom{10}{10}}{2^{10}} = \frac{45 + 10 + 1}{1024} = 0.055,$$

a value nearly **twice** as large and not "significant".

What's wrong? The problem is that we are approximating a highly **discrete** distribution with a continuous one when we refer Z to tables of the $\mathcal{N}(0,1)$ distribution (or, equivalently, Z^2 to tables of the $\chi^2(1)$ distribution). See picture next page!

Solution: Apply the **continuity correction** by moving the numerator of the Z statistic **towards** 0 by 1/2 before squaring:

$$
\begin{aligned}
Z^2 &= \frac{\left(|n_{10} - \frac{1}{2}(n_{10}+n_{01})| - \frac{1}{2}\right)^2}{(n_{10}+n_{01})\frac{1}{2}(1-\frac{1}{2})} \\
&= \frac{(8-5-0.5)^2}{10 \times 0.25} = \frac{2.5^2}{2.5} = 2.5
\end{aligned}
$$

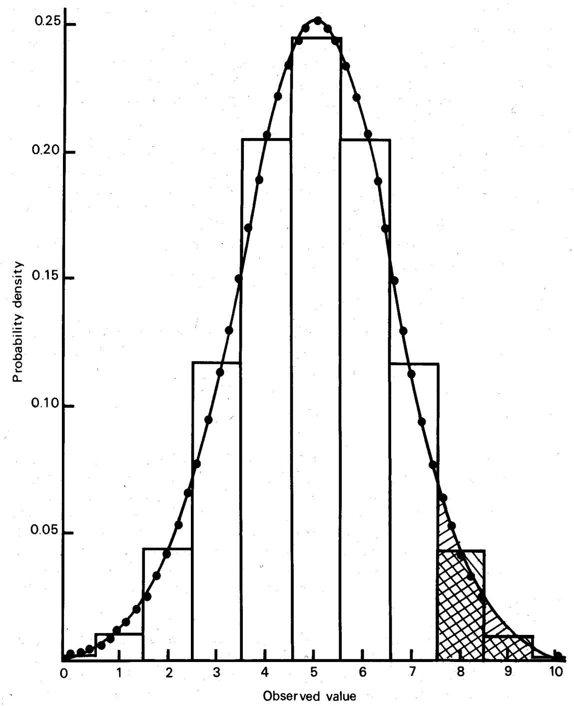
with one-sided $p = 0.057$ (a substantial improvement in accuracy).

132 BRESLOW & DAY

Fig. 4.1 Normal approximation to discrete probability distribution. Note that the discrete probabilities for the values 8, 9 and 10 are better approximated by the area under the normal curve to the right of $7^1/_2$ than by the area under the normal curve to the right of 8.

(hatched) = normal distribution from $7^1/_2$ to infinity

(cross-hatched) = discrete probabilities for the values 8, 9, and 10.



Matched Case-Control Data in STATA

```
. mcci 100 8 2 100

           | Controls             |
Cases      |   Exposed   Unexposed |     Total
-----------+----------------------+----------
   Exposed |       100          8  |       108
 Unexposed |         2        100  |       102
-----------+----------------------+----------
     Total |       102        108  |       210

McNemar's chi2(1) =   3.60     Pr>chi2 = 0.0578
Exact McNemar significance probability = 0.1094

Proportion with factor
  Cases       .5142857
  Controls    .4857143     [95% conf. interval]
              ---------     --------------------
  difference  .0285714     -.0054505   .0625933
  ratio       1.058824       .998108   1.123232
  rel. diff.  .0555556      -.000216   .1113271

  odds ratio         4      .7982264    38.6707 (exact)
```


Fisher's Exact Test for the 2 × 2 Table

Motivation: When a 2×2 table contains cells that have fewer than 5 observations, the normal approximation to the distribution of the log odds ratio (or other summary statistics) is known to be poor. This can lead to incorrect inference since the p-values based on this approximation are not valid.

Solution: Use Fisher's Exact Test.

	D	$\bar{D}$	Total
E			n_1
$\bar{E}$			n_2
Total	m_1	m_2	N

Fisher's Exact Test (cont.)

Example: (Rosner, p. 370) A case-control study was conducted among men aged 50-54 who died from CVD over a 1-month period and similarly aged controls. Dietary habits were ascertained by asking a close relative:

	High Salt	Low Salt	Total
CVD	5	30	35
non-CVD	2	23	25
Total	7	53	60

$$OR = \frac{5 \times 23}{2 \times 30} = 1.92.$$

Interpretation?

Fisher's Exact Test (cont.)

	D	$\bar{D}$	Total
E	a		n_1
$\bar{E}$			n_2
Total	m_1	m_2	N

If we **fix all the margins** then knowledge of the frequency for any one cell of the table will allow the remaining cells to be filled in. Note that $a \geq 0$, $a \leq$ both n_1 and m_1, and a is an integer. Thus there are only a few possible table configurations if either n_1 **or** m_1 is small (with n_1, n_2, m_1, m_2 fixed). For Rosner's CVD example with $n_1 = 25$, $n_2 = 35$, $m_1 = 7$ and m_2=53 the possible tables are

0		25
		35
7	53	60

1		25
		35
7	53	60

2		25
		35
7	53	60

3		25
		35
7	53	60

4		25
		35
7	53	60

5		25
		35
7	53	60

6		25
		35
7	53	60

7		25
		35
7	53	60

Under the null hypothesis, H_0: OR=1, we can use the **hypergeometric distribution** to compute the exact probability of any given table (as specified by a) and thus the exact p-value.

Continuity Corrected χ^2 for 2×2 Table

	High Salt	Low Salt	Total
CVD	5	30	35
non-CVD	2	23	25
Total	7	53	60

Using the hypergeometric distribution we may compute the **exact mean and variance** under H_0: OR=1 as

$$\begin{aligned} E(a|H_0) &= \frac{n_1 m_1}{N} = \frac{25 \times 7}{60} = 2.917 \\ Var(a|H_0) &= \frac{n_1 n_2 m_1 m_2}{(N-1)N^2} = \frac{35 \times 25 \times 7 \times 53}{59 \times 60 \times 60} = 1.528 \end{aligned}$$

Thus the **continuity corrected chi-squared** statistic is

$$\frac{(5 - 2.92 - 0.5)^2}{1.53} = 0.114$$

which may be compared with the **ordinary chi-squared statistic**

$$\frac{60(5 \times 23 - 2 \times 30)^2}{35 \times 25 \times 7 \times 53} = 0.559.$$

The comparison of p-values (one-sided) is:

Exact: 0.37

With continuity correction: 0.37

Without continuity correction: 0.23

STATA

```
cci 5 30 2 23
         |Exp  Unexp | Tot  Prop. Exposed
---------+-----------+----
   Cases |  5    30  |  35      0.1429
Controls |  2    23  |  25      0.0800
---------+-----------+----
   Total |  7    53  |  60      0.1167
                 |                         |
                 |Estimate | [95% Conf. Interval]
                 |-----------------------+------
      Odds ratio |1.917    |  .385 .  (Cornfield)
 Attr. frac. ex. | .478    |-1.595 .  (Cornfield)
                 +-----------------------------
   chi2(1) =     0.56  Pr>chi2 = 0.4546
```


STATA

```
. cci 5 30 2 23, e
         |Exp  Unexp | Tot  Prop. Exposed
---------+-----------+----
   Cases |  5    30  |  35      0.1429
Controls |  2    23  |  25      0.0800
---------+-----------+----
   Total |  7    53  |  60      0.1167
                 |                         |
                 |Estimate | [95% Conf. Interval]
                 |-----------------------+------
      Odds ratio |1.917    |  .385 .  (Cornfield)
 Attr. frac. ex. | .478    |-1.595 .  (Cornfield)
                 +-----------------------------
     1-sided Fisher's exact P = 0.3747
     2-sided Fisher's exact P = 0.6882
```

Data Analysis

"In an analysis, the basic questions to consider are the degree of association between risk for disease and the factors under study, the extent to which the observed associations may result from **bias**, **confounding**, and/or **chance**, and the extent to which they may be described as **causal**." (*Breslow and Day (1980) Vol. I*)

Role of Statistics:

- Quantify systematic variation
  - Treatment effect
  - Dose-response
- Quantify random variation
  - Components of variance of multiple, random sources of variation (multilevel analysis: teacher, class, school)
- Adjust for confounders and selection factors
  - Matched designs (only if necessary)
  - Stratification
  - Regression modeling
- Account in inferences for "chance" (unexplained) variation
  - Standard errors, p-values and confidence intervals

Assessing Causality in Observational Studies

From association to causation according to A. B. Hill (& BDI)

Strength of association

- Helps rule out bias from unmeasured confounders

Consistency of association

- In different population groups
- Using different study designs

Specificity of association

- in disease subcategories
  - specific cell types of cancer
- in exposure groups
  - chest irradiation in prepubertal girls
  - use of filters by cigarette smokers

Relationship in time (exposure precedes effect)

Biological gradient or dose-response

Assessing Causality in Observational Studies

Lack of alternative explanations:

- Adjustment for measured confounders, selection factors
- Sensitivity analyses to guage impact of unmeasured confounders

Coherence of the Evidence: considerations external to the study

- Time trends in population incidence
- Laboratory studies

Initial Treatment of the Data

- Univariate summaries for each variable
- Bivariate tables and scatterplots
  - In case-control studies, two-way tables/plots among **covariates** may be restricted to **controls** to better understand population associations

Grouping of Continuous Variables

Goals:

- Display full range of risk associated with the variable
- Determine form of dose-response relationship

Guidelines:

- A pure non-exposed category should be the baseline level if the numbers appear adequate (e.g., more than 10 controls)
- A simple dichotomy may conceal more information than it reveals
- Use of more than five or six covariate categories will rarely give added insight to the data
- Once the general form of the relationship has been ascertained, the change in risk may be modelled using a functional relationship (linear, quadratic...)

Counfounding

"A confounding variable is a variable that is associated with both the disease and the exposure variable." *Rosner (1995)*

"Confounding is the distortion of a disease/exposure association brought about by the association of other factors with both disease and and exposure, the latter associations with disease being causal." *Breslow & Day (1980)*

"If any factor either increasing or decreasing the risk of a disease besides the characteristic or exposure under study is unequally distributed in the groups that are begin compared with regard to the disease, this itself will give rise to differences in disease frequency in the compared groups. Such distortion, termed confounding, leads to an invalid comparison." *Lilienfeld & Stolley (1994)*

Counfounding

Criteria for a Confounding Factor (direct quote from *Rothman and Greenland, 1998*):

1. A confounding factor must be a risk factor for the disease.
2. A confounding factor must be associated with the exposure under study in the source population (the population at risk from which the cases are derived).
3. A counfounding factor must not be affected by the exposure or the disease. In particular, it cannot be an intermediate step in the causal path between the exposure and the disease.

Causal Diagrams and the Assessment of Confounding

Choosing Confounders for Statistical Adjustment

- Choice should be based on **a priori** considerations
  - Study design/protocol specifies particular "exposure" $\times$ disease association under investigation
  - Confounders selected/measured based on their role as **known risk factors** for the disease
  - See recent work by Pearl, Robins and Greenland for more formal criteria
- Best **not** to select confounders by examination of (internal) study data
  - Selection on basis of statistical significance of association with disease leaves residual confounding effect
  - Selection on basis of resulting change in exposure/disease effect measure destroys opportunity for correct inferences

Choosing Confounders for Statistical Adjustment

- Report results of several planned analyses of primary association:
  - Unadjusted
  - Adjusted for primary set of covariates (known risk factors)
  - Adjusted for primary and secondary set of covariates (known and suspected risk factors)

Positive and Negative Confounding

Positive confounding

Confounder

- Level 1: High risk for disease; High prevalence of exposure
- Level 2: Low risk for disease; Low prevalence of exposure

Negative confounding

Confounder

- Level 1: High risk for disease; Low prevalence of exposure
- Level 2: Low risk for disease; High prevalence of exposure

Positive and Negative Confounding

Extreme example of positive confounding

	C+ E +	C+ E -	C- E +	C- E -	Pooled E +	Pooled E -
Case	90	10	1	9	91	19
Cont	9	1	10	90	19	91
OR	1		1		22.9	

Interpretation of Adjusted Relative Risks

Even in the absence of any association between confounder and exposure in the population, or between covariate and outcome in a randomized treatment trial, the adjusted odds ratio will typically differ from the unadjusted one. Example below due to Mantel (1988):

Exposure	Cases	Noncases	Total	OR	RD
		Stratum 1 – C+			
+	10	90	100		
-	5	95	100	2.11	0.05
		Stratum 2 – C-			
+	95	5	100		
-	90	10	100	2.11	0.05
		Total			
+	105	95	200		
-	95	105	200	1.22	0.05

- Note that exposure/confounder OR $= \frac{100 \times 100}{100 \times 100} = 1$
- Adjusted exposure/disease OR (2.11) represents relative risk **conditional** on fixed values for the covariates (effect of exposure in **subgroup**)
- Overall OR (1.22) represents effect of exposure in

population: average over subgroups (will typically be **smaller**)

- Apparent **paradox** due to **nonlinearity** of OR
  - Risk difference RD not affected by averaging
- Illustrates fundamental point: **meaning of regression coefficients depends on the other variables in the model**

Adjustment of OR via Stratification

- Basic idea of adjusted estimates
  - Compute separate estimate of OR for each stratum
  - Pool OR's rather than data: weighted average of OR's or log(OR)'s with weight proportional to variance
  - Assess homogeneity of ORs across strata by ANOVA or analog
- Basic idea of adjusted test statistic
  - Global null hypothesis: OR in each stratum equals 1
  - Consider Z form of statistic in each stratum: observed value O, expected E and variance V
  - Add up observeds, expecteds and variances separately
  - Form test statistic from summary O, E and V
- Illustration: combination of OR from series of 2×2 tables using
  - Woolf's method
  - Mantel-Haenszel method

Combining Contingency Tables

EXAMPLE: (Rosner 10.8)
A 1985 study identified a group of 518 cancer cases and a group of age- and sex-matched controls by mail questionnaire. The main purpose of the study was to look at the effect of passive smoking on cancer risk. In the study passive smoking was defined as exposure to the cigarette smoke of a spouse who smoked at least one cigarette/day for at least 6 months. One potential *confounding* variable was smoking by the test subjects themselves since personal smoking is related to both cancer risk and having a spouse that smokes. Therefore, it was important to control for personal smoking before looking at the relationship between passive smoking and cancer risk.

[1] - Approach not controlling for personal smoking:

	Passive Smoker		
	Yes	No	
Case	281	228	509
Control	210	279	489
	491	507	998

```
. cci 281 228 210 279

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       281         228  |        509      0.5521
        Controls |       210         279  |        489      0.4294
-----------------+------------------------+----------------------
           Total |       491         507  |        998      0.4920
                 |                        |
                 |      Pt. Est.          |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         1.637406       |  1.275216    2.102466  (Cornfield)
 Attr. frac. ex. |         .3892779       |  .2158191     .524368  (Cornfield)
 Attr. frac. pop |         .2149059       |
                 +-----------------------------------------------
                               chi2(1) =    15.00  Pr>chi2 = 0.0001
```


[2] - Stratified by personal smoking:

<u>Non-smokers</u>

	Passive Smoker		
	Yes	No	
Case	120	111	231
Control	80	155	235
	200	266	466

```
. cci 120 111 80 155

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       120         111  |        231      0.5195
        Controls |        80         155  |        235      0.3404
-----------------+------------------------+----------------------
           Total |       200         266  |        466      0.4292
                 |                        |
                 |      Pt. Est.          |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         2.094595       |  1.442739    3.040946  (Cornfield)
 Attr. frac. ex. |         .5225806       |  .3068739     .671155  (Cornfield)
 Attr. frac. pop |         .2714705       |
                 +-----------------------------------------------
                               chi2(1) =    15.24  Pr>chi2 = 0.0001
```

Smokers

	Passive Smoker		
	Yes	No	
Case	161	117	278
Control	130	124	254
	291	241	532

```
. cci 161 117 130 124

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       161         117  |        278      0.5791
        Controls |       130         124  |        254      0.5118
-----------------+------------------------+----------------------
           Total |       291         241  |        532      0.5470
                 |                        |
                 |      Pt. Est.          |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         1.312558       |  .9324214    1.847675  (Cornfield)
 Attr. frac. ex. |         .2381286       | -.0724765    .4587792  (Cornfield)
 Attr. frac. pop |          .137909       |
                 +-----------------------------------------------
                               chi2(1) =     2.43  Pr>chi2 = 0.1192
```


- We wish to combine the information from both tables to obtain an overall test of significance that takes account of the stratification.

Use the following notation:

	$\boldsymbol{E}$	$\overline{\boldsymbol{E}}$	
$\boldsymbol{D}$	a_i	b_i	$(a_i + b_i)$
$\overline{\boldsymbol{D}}$	c_i	d_i	$(c_i + d_i)$
	$(a_i + c_i)$	$(b_i + d_i)$	N_i

where $i = 1, 2, \ldots, K$ is the number of strata.

We can use the **Mantel-Haenszel Test** to assess the association between disease and exposure after controlling for one or more confounding variables.

Mantel-Haenszel Test

H_o: common odds ratio is 1.0
H_a: common odds ratio $\neq$ 1.0

Procedure:

1. Compute the total observed number in the $(1,1)$ cell over all strata.

$$O = \sum_{i=1}^{K} O_i = \sum_{i=1}^{K} a_i$$

2. Compute the total expected number in the $(1,1)$ cell (under H_o) over all strata.

$$E = \sum_{i=1}^{K} E_i = \sum_{i=1}^{K} \frac{(a_i+b_i)(a_i+c_i)}{N_i}$$

3. Compute the variance, V, of $O-E$.

$$V = \sum_{i=1}^{K} V_i = \sum_{i=1}^{K} \frac{(a_i+b_i)(c_i+d_i)(a_i+c_i)(b_i+d_i)}{N_i^2(N_i-1)}$$

4. Under H_o the test statistic has a $\chi^2(df=1)$ distribution.

$$X^2_{MH} = \frac{(|O-E|-\frac{1}{2})^2}{V}$$

Note: Use the Mantel-Haenszel test only if $V \geq 5$.

EXAMPLE:

$$O = O_1 + O_2 = 120 + 161 = 281$$

$$E_1 = \frac{231 \times 200}{466} = 99.1 \qquad E_2 = \frac{278 \times 291}{532} = 152.1$$

$$E = E_1 + E_2 = 99.1 + 152.1 = 251.2$$

Note: there are more cases who are passive smokers than would be expected based only on their personal smoking habits.

$$V_1 = \frac{231 \times 235 \times 200 \times 266}{466^2 \times 465} = 28.60$$

$$V_2 = \frac{278 \times 254 \times 291 \times 241}{532^2 \times 531} = 32.95$$

$$V = V_1 + V_2 = 28.60 + 32.95 = 61.55$$

Thus,

$$X^2_{MH} = \frac{(|281-251.2|-0.5)^2}{61.55} = \frac{858.49}{61.55} = 13.95$$

p-value $P(\chi^2(1) > 13.95) < 0.001$

Conclude: There is a highly significant association between case-control status and passive smoking exposure even after controlling for personal smoking.

Q: computing with STATA?

```
. infile case passive number smoke using c:/b513/passive.dat
(8 observations read)

. list

          case    passive     number      smoke
  1.         1          1        120          0
  2.         1          0        111          0
  3.         0          1         80          0
  4.         0          0        155          0
  5.         1          1        161          1
  6.         1          0        117          1
  7.         0          1        130          1
  8.         0          0        124          1

. cc case passive [freq=number], by(smoke)

           smoke |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   2.094595     1.442739   3.040946      19.05579 (Cornfield)
               1 |   1.312558     .9324214   1.847675      28.59023 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.637406     1.275216   2.102466               (Cornfield)
    M-H combined |   1.625329     1.263955   2.090024
-----------------+-------------------------------------------------
Test for heterogeneity (M-H)     chi2(1) =     3.270  Pr>chi2 = 0.0706

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     14.42
                                                Pr>chi2 =    0.0001
```


Combining Epidemiologic Measures

Stratum-specific measures may be Attributable Risk, Relative Risk, or Odds Ratio estimates. We will consider only the Odds Ratio.

The common or overall estimates of Attributable Risk and Relative Risk and their corresponding confidence intervals may be found in Kleinbaum/Kupper/Morgenstein, Table 17.16 and Relative Risk in Hennekens and Buring, Table 12.7.

Overview:

Assume there are K 2×2 tables which have been formed by stratification on a (potential) confounding variable.

Use the following notation:

	$\boldsymbol{E}$	$\overline{\boldsymbol{E}}$	
$\boldsymbol{D}$	a_i	b_i	$(a_i + b_i)$
$\overline{\boldsymbol{D}}$	c_i	d_i	$(c_i + d_i)$
	$(a_i + c_i)$	$(b_i + d_i)$	N_i

where $i = 1, 2, \ldots, K$ is the number of strata.

Calculate stratum-specific summaries:

$$\text{odds ratio: } \widehat{\boldsymbol{\Psi}}_i = (a_i d_i)/(b_i c_i)$$
$$\text{log odds ratio: } \widehat{\boldsymbol{\beta}}_i = \log(\widehat{\boldsymbol{\Psi}}_i)$$

$$\text{standard error of log OR: } s_i = \left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right)^{1/2}$$

Under the null hypothesis of no association, $\Psi_i = 1$, or

$\boldsymbol{\beta} = 0$, we have:

$$\frac{\widehat{\boldsymbol{\beta}}_i}{s_i} \approx N(0,1) \quad \rightarrow \quad \left(\frac{\widehat{\boldsymbol{\beta}}_i}{s_i}\right)^2 \approx \chi^2(1)$$
$$\sum_{i=1}^{K} \left(\frac{\widehat{\boldsymbol{\beta}}_i}{s_i}\right)^2 \approx \chi^2(K)$$

Approach:

Partition $\chi^2(K)$ into 2 parts: $\chi^2(K-1)$ and $\chi^2(1)$.

The first part with $K-1$ degrees of freedom tests whether the association is the same in all tables; it is called the **test of homogeneity**, heterogeneity, interaction, or effect modification. (remind you of ANOVA?)

The second part with one degree of freedom tests for association based on information from all of the K tables (Mantel-Haenszel Test).

Method 1: Woolf's Method

If we want to combine estimates, each of which has a different precision (s_i) then it makes sense to "weight" more heavily those estimates that have smaller variance (s_i^2). Statistical theory tells us that the best weights to use are those which are inversely proportional to the variances (know as Gauss-Markov).

<u>Define</u>:

$$w_i = \frac{1}{s_i^2}$$

$$\overline{\boldsymbol{\beta}} = \frac{\sum_{i=1}^{K} w_i \widehat{\boldsymbol{\beta}}_i}{\sum_{i=1}^{K} w_i}$$

Now,

$$\sum_{i=1}^{K} \left(\frac{\widehat{\boldsymbol{\beta}}_i}{s_i} \right)^2 = \sum_{i=1}^{K} w_i (\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})^2 + \sum_{i=1}^{K} w_i (\overline{\boldsymbol{\beta}})^2$$

This is equivalent to:

$$\begin{aligned} \chi^2(K) &= \chi^2(K-1) + \chi^2(1) \\ &= \chi^2_H + \chi^2_A \end{aligned}$$

The first part measures the variation in stratum-specific log

ORs about the average log OR – <u>Test of homogeneity</u>, heterogeneity, interaction, or effect modification.

H_o: $OR_1 = OR_2 = \ldots = OR_K$
H_a: not all are equal (at least one is different)

$$X_H^2 = \sum_{i=1}^{K} w_i (\widehat{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}})^2$$

and we use

$$\text{p-value } = P(\chi^2(K-1) > X_H^2)$$

If the p-value is non-significant (ie. greater than 0.05) then it makes sense to "pool" across the strata either to estimate a single common value and/or to test whether the common value is 0.0 (no association).

Estimate a Common log OR:

Assume that X^2_H is not significant. Then we can use

$$\overline{\beta} = \frac{\sum_{i=1}^K w_i \widehat{\boldsymbol{\beta}}_i}{\sum_{i=1}^K w_i}$$

as a common log OR estimate.

The standard error of this estimate is given by $\overline{V} = (\sum_{i=1}^K w_i)^{-1}$ so that confidence intervals can be obtained as:

$$\overline{\boldsymbol{\beta}} \quad \pm \quad 1.96\sqrt{\overline{V}}$$

$$\exp(\overline{\boldsymbol{\beta}} - 1.96\sqrt{\overline{V}}) \quad , \quad \exp(\overline{\boldsymbol{\beta}} + 1.96\sqrt{\overline{V}})$$

Test for Overall Association:

We can use the confidence interval to test whether the value 0.0 (or 1.0 for OR) lies within the interval. Alternatively, we can use the test statistic X^2_A to test:

H_o: $OR_1 = OR_2 = \ldots = OR_K \ = \ 1.0$
H_a: average OR $\neq 1.0$

$$X^2_A = \sum_{i=1}^K w_i \overline{\boldsymbol{\beta}}^2$$

which would be compared to a $\chi^2(1)$.

$$\text{p-value} \ = P(\chi^2(1) > X^2_A)$$

Q: How did we answer this type of question with linear regression?

Woolf's Method with Passive Smoking Data

a b / c d	$\widehat{\text{OR}} = \frac{ab}{cd}$	$\hat{\beta}$	$\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)$
120 111 / 80 155	2.095	0.7394	0.03629
161 117 / 130 124	1.313	0.2720	0.03051

Combined OR estimate

$$\begin{aligned}
\hat{\beta}_W = \bar{\beta} &= \frac{\frac{1}{.03629} \times 0.7394 + \frac{1}{0.03051} \times 0.2720}{\frac{1}{.03629} + \frac{1}{0.03051}} \\
&= \frac{27.55 \times 0.7394 + 32.77 \times 0.2720}{27.55 + 32.77} \\
&= 0.4855 \\
\widehat{\text{OR}} &= \exp(0.4855) = 1.625
\end{aligned}$$

Confidence Interval based on $\hat{\beta}$

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\beta}) &= \frac{1}{27.55 + 32.77} = 0.01658 \\
(\beta_{\text{L}}, \beta_{\text{U}}) &= 0.4855 \pm 1.96 \times \sqrt{0.01658} \\
&= (0.2331, 0.7379) \\
(\text{OR}_{\text{L}}, \text{OR}_{\text{U}}) &= (\exp^{.2331}, \exp^{.7379}) = (1.263, 2.091)
\end{aligned}$$

Example

Test of Association

$$Z^2 = \frac{\hat{\beta}^2}{\widehat{\text{Var}}(\hat{\beta})} = \frac{.4855^2}{.01658} = 14.216, \quad p = 0.0000$$

Test of Homogeneity

$$\begin{aligned} X^2_{\text{HOM}} &= 27.55(0.7394)^2 + 32.77(0.2720)^2 - 14.216 \\ &= 3.270 \sim \chi^2_1, \ \ p = 0.07 \end{aligned}$$

Problem: Method breaks down with "sparse data" (small cell entries)

- Can't take logs from 0 cell entries
- Addition of 1/2 biases OR estimate

Method 2: Mantel-Haenszel Method

Recall we were able to test the hypothesis

H_o: $OR_1 = OR_2 = \ldots = OR_K \ = \ 1.0$
H_a: average OR $\neq 1.0$

using the Mantel-Haenszel statistic

$$X^2_{MH} = \frac{(|O-E| - \frac{1}{2})^2}{V}$$

which has a $\chi^2(1)$ distribution under H_o.

Mantel-Haenszel Estimate

Instead of combining the logORs using weight (Woolf), MH combines the ORs directly using weights that approximate the variance of the stratum-specific estimates (when $\Psi \approx 1$).

$$\begin{aligned} \widehat{\boldsymbol{\Psi}}_{MH} &= \frac{\sum_{i=1}^K w_i \widehat{\boldsymbol{\Psi}}_i}{\sum_{i=1}^K w_i} \\ w_i &= (b_i c_i)/N_i \\ \widehat{\boldsymbol{\Psi}}_{MH} &= \frac{\sum_{i=1}^K a_i d_i/N_i}{\sum_{i=1}^K b_i c_i/N_i} \end{aligned}$$

Confidence Interval

- Robins, Breslow, and Greenland (1986)
- Test Based method (Miettinen 1976)

Mantel-Haenszel and Confidence Intervals

We have introduced:

- The Mantel-Haenszel Estimate, $\widehat{\boldsymbol{\Psi}}_{MH}$.
- The Mantel-Haenszel Test, $X^2_{MH} \sim \chi^2(1)$.

Q: How can we compute a confidence interval for the MH estimate?

A: Using the estimate and the test we can construct a test-based (Miettinen, 1976) confidence interval:

$$\begin{aligned} \log \boldsymbol{\Psi} &\in \log \widehat{\boldsymbol{\Psi}}_{MH} \pm Q_Z^{1-\alpha/2} \left(\frac{\log \widehat{\boldsymbol{\Psi}}_{MH}}{\sqrt{X^2_{MH}}} \right) \\ \Psi &\in \widehat{\boldsymbol{\Psi}}_{MH}^{(1 \pm Q_Z^{1-\alpha/2}/\sqrt{X^2_{MH}})} \end{aligned}$$

Note: only recommended if $\widehat{\boldsymbol{\Psi}}_{MH}$ isn't too far from 1 – ie. between 1/5 and 5.

Example:

In out analysis of Cancer and Passive Smoking we stratified on Personal Smoking and obtained:

$$\widehat{\boldsymbol{\Psi}}_{MH} = 1.63$$

$$X^2_{MH} = 14.42$$

from which we could obtain a CI:

$$\begin{aligned}
\log \boldsymbol{\Psi} &\in \log \widehat{\boldsymbol{\Psi}}_{MH} \pm Q_Z^{1-\alpha/2} \left(\frac{\log \widehat{\boldsymbol{\Psi}}_{MH}}{\sqrt{X^2_{MH}}} \right) \\
& \quad 0.489 \pm 1.96(0.489/\sqrt{14.42}) \\
& \quad 0.489 \pm 0.252 \\
\log \boldsymbol{\Psi} &\in (\ 0.237,\ 0.741\ ) \\
\Psi &\in (\ \exp(0.237),\ \exp(0.741)\ ) \\
&\in (\ 1.267,\ 2.098\ )
\end{aligned}$$

with Robins-Greenland-Breslow Variance

Define (for each 2×2 table):

$$R_i = \frac{a_i \times d_i}{N_i} \qquad S_i = \frac{b_i \times c_i}{N_i}$$

$$P_i = \frac{a_i + d_i}{N_i} \qquad Q_i = \frac{b_i + c_i}{N_i}$$

Then MH estimate and RGB variance given by

$$\begin{aligned}
\hat{\psi}_{\text{MH}} &= \frac{\sum_i R_i}{\sum_i S_i} \equiv \frac{R_+}{S_+} \\
\widehat{\text{Var}}(\log \hat{\psi}_{\text{MH}}) &= \frac{\sum_i P_i R_i}{2R_+^2} + \frac{\sum_i (P_i S_i + Q_i R_i)}{2R_+ S_+} + \frac{\sum_i Q_i S_i}{2S_+^2}
\end{aligned}$$

Why does it work? (estimating equation theory)

- $\text{E}(R_i - \psi_i S_i) = 0 \Rightarrow R_+/S_+$ "unbiased"
- $\text{Var}(R_i - \psi_i S_i) = \frac{1}{2}\text{E}\{(R_i + \psi_i S_i)(P_i + \psi_i Q_i)\}$

Example: MH Estimation with Ille-et-Vilaine Data

- Exposure = 80g + alcohol per day
- Stratified by age (six groups)
- See next page for calculations (also B&D I, p. 145)

Mantel-Haenszel Estimate

table (a b / c d)	$R = ad/N$	$S = bc/N$	$P = (a+b)/N$	$Q = (b+c)/N$
1, 0 / 9, 106	0.914	0.000	0.922	0.078
4, 5 / 26, 164	3.296	0.653	0.844	0.156
25, 21 / 29, 138	16.197	2.859	0.765	0.235
42, 34 / 27, 139	24.124	3.793	0.748	0.252
19, 36 / 18, 88	10.385	4.025	0.665	0.335
5, 8 / 0, 31	3.523	0.000	0.818	0.182
Totals	R_+=58.439	S_+ =11.330		

$\sum PR = 43.848$ $\sum PS = 8.252$ $\sum QR = 14.591$ $\sum QS = 3.079$

Mantel-Haenszel Estimate

Summing up

$$
\begin{aligned}
\hat{\psi}_{\mathrm{MH}} &= 58.439/11.330 = 5.158 \\
\mathrm{Var}(\log \hat{\psi}_{\mathrm{MH}}) &= \frac{43.848}{2 \cdot 58.439^2} + \frac{8.252 + 14.591}{2 \cdot 58.439 \cdot 11.330} + \frac{3.079}{2 \cdot 11.330^2} \\
&= 0.03566 \\
(\psi_{\mathrm{L}}, \psi_{\mathrm{U}}) &= \exp\left[\log(5.158) \pm 1.96 \times \sqrt{0.03566}\right] = (3.56, 7.47)
\end{aligned}
$$

Compare with Woolf estimate and CI (add 0.5 to each cell)

$$\hat{\beta}_{\mathrm{W}} = 1.6141 \pm 0.1881 \quad \hat{\psi}_{\mathrm{W}} = \exp(1.6141) = 5.023$$

$$(\psi_{\mathrm{L}}, \psi_{\mathrm{U}}) = \exp\left[1.6141 \pm 1.96 \times 0.1881\right] = (3.47, 7.26)$$

Q: Is there a **test for homogeneity** that can be used with MH?

A: Yes, the “Breslow-Day” test.

The "Breslow-Day" Test for Homogeneity

Define :

- $A_i(\psi)$ = "expected value" for a_i in a table with fixed marginal totals $(n_{1i}, n_{2i}, m_{1i}, m_{21})$ when the odds ratio is known to equal ψ, by solving the **quadratic** (in A_i) equation:

$$A_i \times (n_{2i} - m_{1i} + A_i) = \psi \times (n_{1i} - A_i) \times (m_{1i} - A_i)$$

- $\text{Var}(a_i | n_{1i}, n_{2i}, m_{1i}, m_{2i}; \psi)$, corresponding variance,

$$\left\{ \frac{1}{A_i(\psi)} + \frac{1}{n_{21} - m_{1i} + A_i(\psi)} + \frac{1}{n_{1i} - A_i(\psi)} + \frac{1}{m_{1i} - A_i(\psi)} \right\}^{-1}$$

Then the **corrected** (by Tarone) "Breslow-Day" test is

$$\sum_{i=1}^{I} \frac{\left[a_i - A_i(\hat{\psi}_{\text{MH}})\right]^2}{\text{Var}(a_i; \hat{\psi}_{\text{MH}})} - \frac{\left[\sum_{i=1}^{I} a_i - \sum_{i=1}^{I} A_i(\hat{\psi}_{\text{MH}})\right]^2}{\sum_{i=1} \text{Var}(a_i; \hat{\psi}_{\text{MH}}}$$

- Often the correction term is near 0, reflecting the **efficiency** of $\hat{\psi}_{\text{MH}}$
- In contradistinction to $\hat{\psi}_{\text{MH}}$ and the RGB variance, the B-D test should only be used with a "small" number of "large" tables
- See example next page

Mantel-Haenszel and Homogeneity

Age (yrs)	a | b / c | d	$A(\hat{\psi}_{\text{MH}})$	$\text{Var}(a; \hat{\psi}_{\text{MH}})$
25-34	1 | 0 / 9 | 106	0.328	0.215
35-44	4 | 5 / 26 | 164	4.104	2.030
45-54	25 | 21 / 29 | 138	24.500	7.782
55-64	42 | 34 / 27 | 139	40.135	10.557
65-74	19 | 36 / 18 | 88	23.740	6.238
75+	5 | 8 / 0 | 31	3.203	0.996
Totals	96 | 104 / 109 | 666	96.000	27.819

$$
\begin{aligned}
\chi^2_{BD}(5) &= \frac{(1-0.328)^2}{0.215} + \cdots + \frac{(5-3.203)^2}{0.996} - \frac{(96-96.000)^2}{27.819} \\
&= 9.32, \quad p = 0.10
\end{aligned}
$$

MH Estimation in sc stata

- Using `cc`

```
. use ille-et-vilaine
. gen alcexp=alc
. recode alcexp 1/2=0 3/4=1
. cc cc alcexp [freq=freq], by(age) bd

   age |     OR    [95% Conf. Interval]  M-H Weight
-------+-------------------------------------------
 25-34 |         .          0          .          0
 35-44 | 5.046154    1.37344   18.64205    .6532663
 45-54 | 5.665025    2.81505   11.40813    2.859155
 55-64 | 6.359477    3.45929   11.6933     3.793388
 65-74 | 2.580247    1.22547    5.43615    4.024845
   75+ |         .    4.38873          .          0
-------+-------------------------------------------
 Crude | 5.640085     4.003217   7.94673
 M-H d | 5.157623     3.562131   7.46774
----------------+---------------------------------
Test of homog (M-H) chi2(3) = 3.78  Pr>chi2 = 0.2868
Test of homog (B-D) chi2(5) = 9.32  Pr>chi2 = 0.0968

Test that combined OR = 1:
Mantel-Haenszel chi2(1) =  85.01 Pr>chi2 =    0.0000
```

Mantel-Haenszel Estimate

- Using `mhodds` by Clayton and Hills

```
. mhodds cc alcexp age [freq=freq]

Mantel-Haenszel estimate of the odds ratio
Comparing alcexp==1 vs alcexp==0, controlling for age

Odds ratio  chi2(1)  P>chi2     [95% Conf. Interval]
  5.157623    85.01  0.0000      3.494918   7.611359
----------------------------------------------------
```


Mantel-Haenszel Estimate

```
. mhodds cc alcexp [freq=freq], by(age)

Maximum likelihood estimate of the odds ratio
Comparing alcexp==1 vs alcexp==0, by age
  age | Odds ratio chi2(1) P>chi2  [95% Conf. Interval]
------+-------------------------------------------------------
25-34 |          .   10.60 0.0011         .         .
35-44 |   5.046154    6.32 0.0119  1.238891  20.55361
45-54 |   5.665025   25.94 0.0000  2.664715  12.04350
55-64 |   6.359477   38.74 0.0000  3.256609  12.41873
65-74 |   2.580247    6.27 0.0123  1.194820  5.572115
  75+ |          .   13.15 0.0003         .         .
------+-------------------------------------------------------

Odds ratio  chi2(1) P>chi2     [95% Conf. Interval]
 5.157623    85.01  0.0000      3.494918   7.611359
----------------------------------------------------
Test of homogeneity of ORs (approx):
chi2(5)   =    8.28   Pr>chi2   =  0.1416
```

Mantel-Haenszel and Paired Data

Q: What do we obtain if we use Mantel-Haenszel to obtain a common odds ratio across strata obtained from pair-matching?

A: The only possible tables are:

	E	$\overline{E}$	
D			1
$\overline{D}$			1
			2

$a_i \times d_i \ =$

$b_i \times c_i \ =$

	E	$\overline{E}$	
D			1
$\overline{D}$			1
			2

$a_i \times d_i \ =$

$b_i \times c_i \ =$

	E	$\overline{E}$	
D			1
$\overline{D}$			1
			2

$a_i \times d_i \ =$

$b_i \times c_i \ =$

	E	$\overline{E}$	
D			1
$\overline{D}$			1
			2

$a_i \times d_i \ =$

$b_i \times c_i \ =$

Mantel-Haenszel Estimate

And the Mantel-Haenszel odds ratio estimate is given by:

$$\hat{\psi}_{\text{MH}} = \frac{\sum_i a_i d_i / 2}{\sum_i b_i c_i / 2} = \frac{n_{01}}{n_{10}}$$

where

$$n_{01} = \# \text{ of pairs with } E = 1 \text{ for } D, \text{ and } E = 0 \text{ for } \overline{D}.$$

$$n_{10} = \# \text{ of pairs with } E = 0 \text{ for } D, \text{ and } E = 1 \text{ for } \overline{D}.$$

Summary

- The Woolf Method works best if the tables are not too small (N_i). It is recommended when there are a few to moderate number of tables (K).

- The Mantel-Haenszel Method works best when the number of tables (K) is not too small and when there are a large number of small tables (N_i).

- For most situations there is little difference between the two methods; the largest difference is usually seen in the confidence intervals. For tables containing cells with 0's the Mantel-Haenszel procedure is best; Woolf's procedure requires zero-cell correction (add 1/2 to each cell).

Procedure :

1. Compute stratum-specific measures.
2. Evaluate stratum-specific estimates by a test of homogeneity. Consider test results in light of sample size.
3. If the homogeneity test result is non-significant then we're happy to consider a common estimate, pooling across all strata, then
   (a) calculate an overall (common) summary measure
   (b) test for significant association
   (c) calculate confidence interval
4. If the homogeneity test result is significant then we are concerned that the ORs vary across strata. We may
   (a) If the direction of association ($\pm$) is same and the difference is small in magnitude, then we may proceed as in 3 above (calculating average summary), but report on the test of homogeneity.

   (b) If the direction of the association is different, then
   - report results from test of homogeneity
   - report stratum-specific measures and confidence intervals.
   - does the average make sense at all?

Limitations of χ^2 with Ordered Categories

Example: Smoking and quality of life:

Self Report Quality of Health

Smoke	Poor	Fair	Good	V.Good	Exc..	Total
No	11	27	42	53	11	144
Yes	7	15	16	13	1	52
	18	42	58	66	12	196

- Is there an association between the self report of health and smoking?
  - Pearson chi-squared statistic gives:
    $$\sum(O-E)^2/E = 6.881 \sim \chi^2_{(2-1)(5-1)} = \chi^2_4$$
    with p-value=0.142.
  - Does that imply that there is no **detectable** association?
  - Does the Pearson statistic change value if the rows are permuted? if the columns are permuted? Should it?
- Is there an **alternative** for the Pearson statistic?

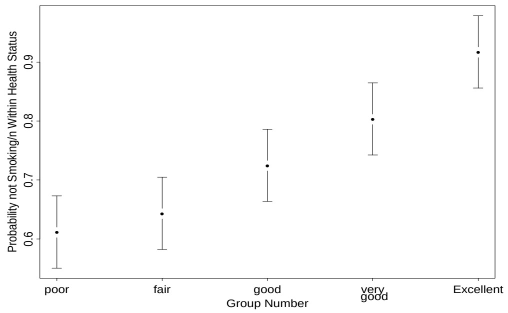
An Alternative View of the Data

Smoke	Poor	Fair	Good	V.Good	Exc.	Total
No	11	27	42	53	11	144
Yes	7	15	16	13	1	52
n_k	18	42	58	66	12	N =196
$\hat{p}_k$	11/18	27/42	42/58	53/66	11/12	$\bar{p} = \frac{144}{196}$
	0.61	0.64	0.72	0.80	0.92	0.73

- Global chi-squared statistic tests

$$\begin{aligned} \mathrm{H}_0 &: p_1 = p_2 = \cdots = p_K \text{ vs} \\ \mathrm{H}_\mathrm{A} &: p_k \neq p_j \text{ for some } k, j \end{aligned}$$

which ignores the **pattern**:



Test for Trend in Binomial Proportions

Q: Can we make the hypothesis more specific?

A: Yes, use as the alternative

$$\mathrm{H}_A : p_1 \leq p_2 \leq p_3 \cdots p_K \text{ or } \mathrm{H}_A : p_1 \geq p_2 \geq p_3 \cdots p_K$$

with inequality for at least one pair.

Define: (See Rosner or B&D I, §4.5 for more detail)

$$A = \overbrace{\sum n_k(\hat{p}_k - \bar{p})(x_k - \bar{x})}^{(N-1)\mathrm{Cov}(x_k, \hat{p}_k)}$$

$$B = \underbrace{\bar{p}(1-\bar{p})}_{\mathrm{Var}(\mathrm{bin}(1,\bar{p}))} \underbrace{\left\{ \left(\sum n_k x_k^2\right) - \left(\sum n_k x_k\right)^2 / N \right\}}_{(N-1)\mathrm{Var}(x)}$$

$$\bar{x} = \sum n_k x_k / N \text{ and } \bar{p} = n_1/N$$

$$\chi_1^2 = A^2/B$$

where x_k is the (transform of) "dose" associated with k^{th} level.

"Doses" must be selected **a priori** but may be

- Equally spaced (default in most programs – makes no difference how centered or scaled)
- Multiplicative: x = 1, 2, 4, 8, 16

- Logarithmic: $x = 0, \log(2), \log(3), \log(4)$

Correction for continuity (linear trend only):

- Reduce the absolute value of A by **half the distance** between neighboring x_k
- Multiply the resulting χ^2_1 by $(N-1)/N$

Known as : Cochran-Armitage test for trend

Testing for Trend in STATA

```
. input smok freq1 freq2 freq3 freq4 freq5
  1. 0 11 27 42 53 11
  2. 1  7 15 16 13  1
  3. end

. reshape long freq, i(smok) j(health)
(note:  j = 1 2 3 4 5)
. tabodds smok health [freq=freq]

--------+--------------------------------------
health  |cases  conts    odds [95% Conf. Interval]
--------+--------------------------------------
     1  |    7     11  0.63636   0.24669   1.64156
     2  |   15     27  0.55556   0.29554   1.04434
     3  |   16     42  0.38095   0.21419   0.67755
     4  |   13     53  0.24528   0.13373   0.44990
     5  |    1     11  0.09091   0.01174   0.70414
------------+----------------------------------
Test of homogeneity:  chi2(4)  =     6.85
  (equal odds)        Pr>chi2  =   0.1443

Score test for trend: chi2(1)  =     6.63
                      Pr>chi2  =   0.0100
```

Brief Review

- $R \times C$ contingency table
  - Test for homogeneity (Pearson chi-squared)
  - Test for trend (Cochran-Armitage)
- Single 2×2 table
  - Different sampling schemes
    1. Cohort (row totals fixed)
    2. Case-control (column totals fixed)
    3. Cross-sectional (grand total fixed)
  - Different measures of association

    RD (Designs 1 & 3)
    RR (Designs 1 & 3)
    OR (Designs 1, 2 & 3)
  - Test of association

    Pearson chi-squared
    Fisher exact

Brief Review (cont.)

- Series of 2×2 tables
  - Summary (combined) OR estimate

    Woolf weighted average ($\bar{\beta}$) of $\log \hat{\mathrm{OR}}$
    Mantel-Haenszel better, esp. for "sparse" data
  - Test for Association: OR = 1 vs OR constant, $\neq 1$

    Woolf Z^2 based on Woolf estimate $\bar{\beta}$
    Mantel-Haenszel test (with continuity correction)
  - Test for Homogeneity (Interaction, Effect Modification)

    Woolf ANOVA analog $\sum_i w_i(\hat{\beta}_i - \bar{\beta})^2$
    Corrected Breslow-Day "score" test
  - Paired binary data as extreme case of stratification of 2×2 table