- Week 3 –

Some Common Measures of Association for 2x2 Tables

Professor Eric Beh

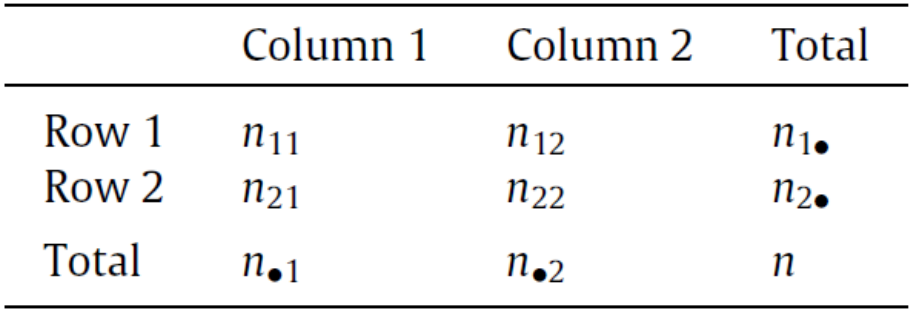
School of Mathematical & Physical Sciences
University of Newcastle

Pearson’s Statistic for a 2x2 Table

In last weeks lectures we discussed, in part, the history, development and issues concerned with Pearson’s chi-squared statistic.

Here we shall adopt the following notation for a single 2x2 contingency table:

Notation for a 2 × 2 contingency table.



	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

For a single 2x2 table, Pearson’s chi-squared statistic is

$$X^2 = n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}}$$

Alternatively, the statistic can be simply expressed as

$$X^2 = n\left(\frac{P_1 - p_{\bullet 1}}{p_{2\bullet}}\right)^2\left(\frac{p_{1\bullet}p_{2\bullet}}{p_{\bullet 1}p_{\bullet 2}}\right)$$

$$= n\frac{\left(p_{11} - p_{1\bullet}p_{\bullet 1}\right)^2}{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}$$

where $P_1 = p_{11} / p_{1\bullet}$ is the conditional probability of an individual/unit being classified into "Column 1" given that they are classified in "Row 1".

Note that

$$X^2 = n\tilde{r}_{11}^2$$

where

$$\tilde{r}_{11} = \frac{p_{11} - p_{1\bullet}p_{\bullet 1}}{\sqrt{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}} = \frac{p_{11} - p_{1\bullet}p_{\bullet 1}}{\sqrt{p_{1\bullet}p_{\bullet 1}\left(1 - p_{1\bullet}\right)\left(1 - p_{\bullet 1}\right)}}$$

is termed the *adjusted standardised residual* of the (1, 1)th cell and is asymptotically standard normally distributed (see Week 4 notes for more)

Example

Occupational Exposure (yrs)	Asbestosis		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

Beh and Smith (2011)

```
> asbestos.dat
      No Yes
0-19 522 203
20+   53 339
> asbestos.dat/1117
             No         Yes
0-19 0.46732319 0.1817368
20+  0.04744852 0.3034915
>
> apply(asbestos.dat/1117,1,sum)
   0-19     20+
0.64906 0.35094
>
> apply(asbestos.dat/1117,2,sum)
       No         Yes
0.5147717 0.4852283
>
```

$$X^2 = n\frac{(p_{11} - p_{1\bullet}p_{\bullet 1})^2}{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}$$

Alternatively, in R

```
> asbestos.dat
      No Yes
0-19 522 203
20+   53 339
>
> chisq.test(asbestos.dat, correct=F)

        Pearson's Chi-squared test

data:  asbestos.dat
X-squared = 348.3524, df = 1, p-value < 2.2e-16

>
```

Here, `correct = F` has been specified to ensure Yates’ continuity correction (see next slides) is not imposed – by default the correction is incorporated into the analysis of a 2x2 contingency table.

Yates' Continuity Correction

Frank Yates
(1902 - 1994)

Yates (1934) argued that the chi-squared statistic with 1 degree of freedom gives p-values from Pearson's statistic that typically underestimate the true p-values. As a result he proposed an adjustment to Pearson's chi-squared statistic of a 2x2 contingency table:

$$X^2 = n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(\left|p_{ij} - p_{i\bullet}p_{\bullet j}\right| - \frac{1}{2n}\right)^2}{p_{i\bullet}p_{\bullet j}}$$

or, equivalently

$$X^2 = n\frac{\left(\left|p_{11} - p_{1\bullet}p_{\bullet 1}\right| - 0.5/n\right)^2}{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}$$

and are both a chi-squared random variable with 1 degree of freedom.

Example

Occupational Exposure (yrs)	Asbestosis		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

```
> asbestos.dat
      No Yes
0-19 522 203
20+   53 339
>
> chisq.test(asbestos.dat)

        Pearson's Chi-squared test with Yates' continuity
correction

data:  asbestos.dat
X-squared = 346.0151, df = 1, p-value < 2.2e-16

>
```

or . . .

$$X^2 = n\frac{\left(\left|p_{11} - p_{1\bullet}p_{\bullet 1}\right| - 0.5/n\right)^2}{p_{1\bullet}p_{2\bullet}p_{\bullet 1}p_{\bullet 2}}$$

Fleiss, Levin & Paik (2003, Chapter 3) provide an excellent review of strategies for including Yate's continuity correction. However, studies have revealed that incorporating the correction is not essential. Mantel and Greenhouse (1968, pg 30) concede that while they prefer to use the continuity correction, "this is not to say that pathologies may not occur with its use".

In fact, Agresti (2002, pg 103) points out that, due to advances in software development, Yates' continuity correction is no longer needed.

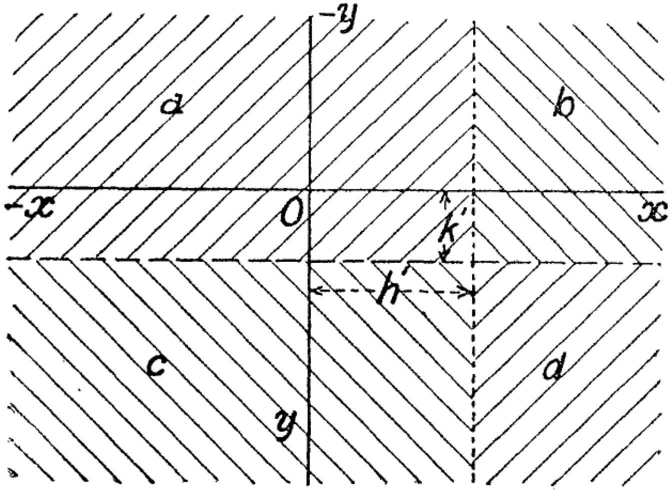
Stuart, Ord & Arnold (2002, pg 413) say on the topic

> *". . . we tentatively conclude that using X^2 without the correction is to be preferred, but since the controversy has continued for almost 100 years, we hesitate to claim a final statement."*

Pearson’s Tetrachoric Correlation

Pearson (1900) formally developed what we now know as the correlation coefficient.

Pearson (1900) considered two standard normally distributed random variables X ~ N(0, 1) and Y ~ N(0, 1). By denoting the correlation between X and Y by r, we dichotomise the variables such that. . .



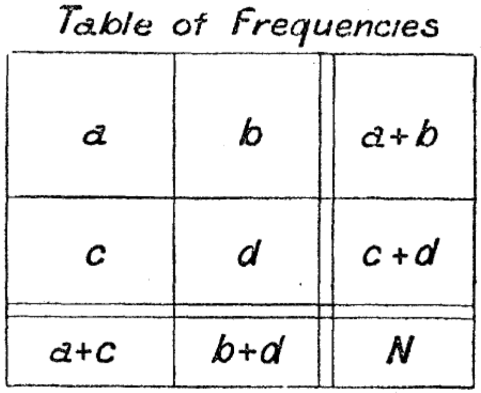


Table of Frequencies

a	b	a+b
c	d	c+d
a+c	b+d	N

Then

$$d = \frac{N}{2\pi\sqrt{1-\underset{?}{\boxed{r^2}}}}\int_{h'}^{\infty}\int_{k'}^{\infty}\exp\left(-\frac{x^2-2rxy+y^2}{2\left(1-\underset{?}{\boxed{r^2}}\right)}\right)dxdy$$

In order to determine the correlation, r, Pearson resorted to a tetrachoric series approximation for above double integral and hence, r is referred to as the tetrachoric correlation.

$$d = \frac{1}{2\pi\sqrt{1-\underset{?}{\boxed{r^2}}}}\int_{h'}^{\infty}\int_{k'}^{\infty}\exp\left(-\frac{x^2-2rxy+y^2}{2\left(1-\underset{?}{\boxed{r^2}}\right)}\right)dxdy$$

The calculation of r is computationally complex – although one may consider Froemel (1971), Martinson and Hamdan (1975), Brown (1977) and Brown and Benedetti (1977). Fleming (2005) provides an improved algorithm which he refers to as TETCORR which may be freely downloaded from

http://swppr.com/TETCORR.htm

In R, one may consider the function `tetrachoric` in the `psych` library.

Approximations of the Tetrachoric Correlation

To overcome the computational difficulties that he faced at the time, Pearson (1900, pg 7) proposed a number of approximations to his correlation. These include

$$r_1 = \sin\left[2\pi\left(\frac{n_{11}n_{22} - n_{12}n_{21}}{n^2}\right)\right] \qquad r_2 = \cos\left[\pi\left(\frac{n_{12}}{n_{1\bullet}}\right)\right]$$

$$r_3 = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}}}$$

Other approximations are

$$Q_1 = \sin\left(\frac{\pi}{2}\frac{n_{11}n_{22} - n_{12}n_{21}}{n_{1\bullet}n_{2\bullet}}\right) \quad \text{when} \quad n_{11}n_{22} > n_{12}n_{21}$$

$$Q_2 = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

. . . and is called Yule’s (1900) Q.

$$Q_3 = \sin\left(\frac{\pi}{2}\frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}}\right)$$

Two more approximations of Pearson's tetrachoric correlation include

$$Q_4 = \sin\left(\frac{\pi}{2}\left(1 + \frac{2nn_{12}n_{21}}{(n_{12}+n_{21})(n_{11}n_{22}-n_{12}n_{21})}\right)^{-1}\right) \quad \text{when} \quad n_{11}n_{22} > n_{12}n_{21}$$

and

$$Q_5 = \sin\left(\frac{\pi}{2}\frac{1}{\sqrt{1+\kappa^2}}\right)$$

where

$$\kappa^2 = \frac{4n^2 n_{11}n_{22}n_{12}n_{21}}{(n_{11}+n_{22})(n_{12}+n_{21})(n_{11}n_{22}-n_{12}n_{21})^2}$$

Note:

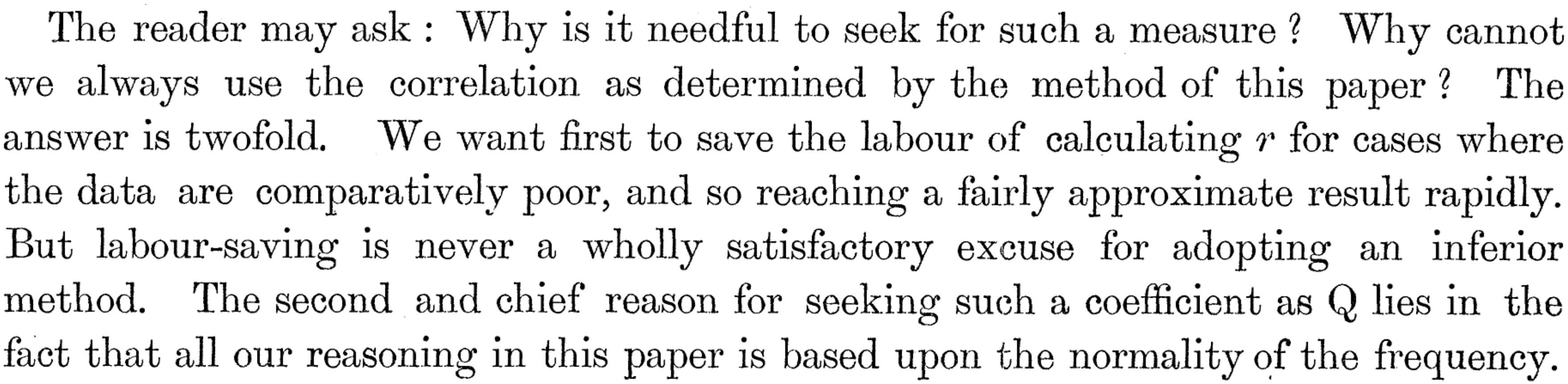
The commonly used measures of correlation between two dichotomous variables cross-classified is r_3 and Q_2.

With so many approximations of his tetrachoric correlation, which did Pearson (1900) prefer?

- He felt Q_1 was of "little service"
- He felt that Q_2, Q_3, Q_4 and Q_5 were satisfactory
- Based on the data he analysed, he preferred Q_2

Pearson's recommendation for Q_2 is based on empirical evidence of its performance for approximating r, but it worth considering that at the turn of the century Yule (a categorical data analyst of great repute later) was also protégé of Karl Pearson.

Pearson (1900, pg 17) commented



The reader may ask: Why is it needful to seek for such a measure? Why cannot we always use the correlation as determined by the method of this paper? The answer is twofold. We want first to save the labour of calculating r for cases where the data are comparatively poor, and so reaching a fairly approximate result rapidly. But labour-saving is never a wholly satisfactory excuse for adopting an inferior method. The second and chief reason for seeking such a coefficient as Q lies in the fact that all our reasoning in this paper is based upon the normality of the frequency.

However . . .

While Yule was Pearson's protégé, the Pearson condition that his correlation is

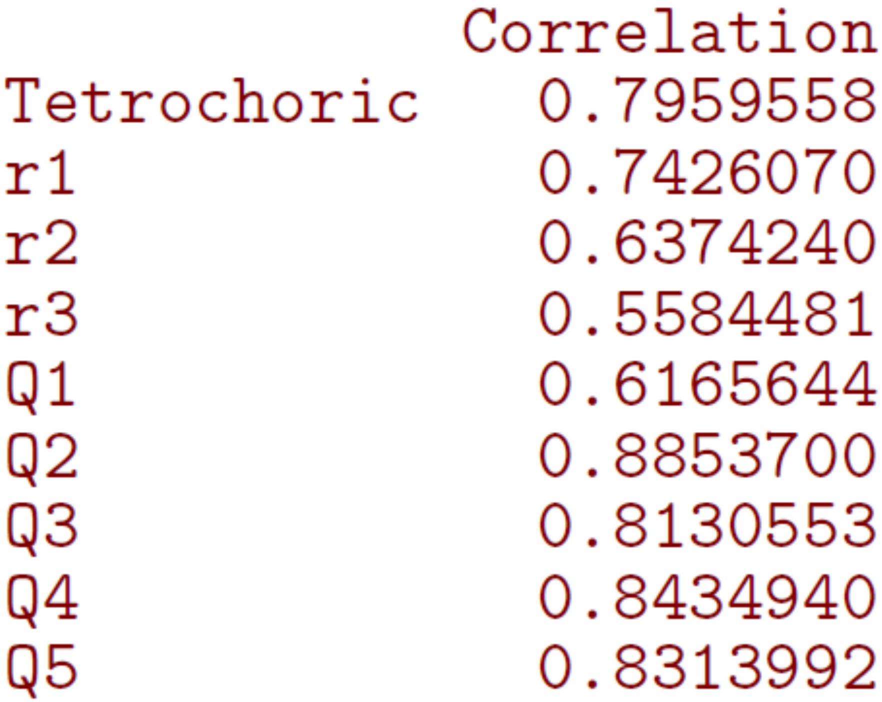
"based upon the normality of the frequency"

would cause their collaboration and friendship to end very, very sourly.

Example

Occupational Exposure (yrs)	Asbestosis		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

Selikoff's asbestos data



```
              Correlation
Tetrachoric     0.7959558
r1              0.7426070
r2              0.6374240
r3              0.5584481
Q1              0.6165644
Q2              0.8853700
Q3              0.8130553
Q4              0.8434940
Q5              0.8313992
```

Using `tetrachoric` in the `psych` library

"little service"

satisfactory

The Odds Ratio

One of the most simple, influential and diversely used measures of association for a 2x2 table is the odds ratio.

$$\theta = \frac{\boxed{n_{11}n_{22}}}{\boxed{n_{12}n_{21}}}$$

	Column 1	Column 2	Total
Row 1	n_{11}	n_{12}	$n_{1\bullet}$
Row 2	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Since the odds ratio is just the ratio of the product of the elements of the main diagonal of the 2x2 table with the product of the elements of the off diagonal, θ has also come to be known as the *cross-product ratio*.

The origins of this odds ratio date back to Jerome Cornfield (1951).

In his investigation of the link between smoking and lung cancer Cornfield was interested in the relative proportion of those smokers who developed cancer, and those who didn't.

Jerome Cornfield
(1912–1979)

He also undertook a rigorous review of its properties, which include:

1. Since no cell proportion can be negative, θ *cannot be negative*.
2. $\theta = 1$: There is no association between the two variables
3. $\theta > 1$: There is a positive association between the two variables
4. $0 < \theta < 1$: There is a negative association between the two variables

Example

Occupational Exposure (yrs)	Asbestosis		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

For those with an **asbestos exposure of less than 20 years** the odds of contracting asbestosis can be calculated as: 203/522 = 0.39.

Since this result is less than 1, it suggest that individuals exposed to asbestos for less than 20 years are not likely to be diagnosed with asbestosis.

For individuals **exposed to asbestos for 20 years or more**, the odds that they are then diagnosed with asbestosis is: 339/53 = 6.396.

That is, workers exposed for a relatively long period of time are around 6.4 times more likely to be diagnosed with asbestosis when compared to their colleagues with shorter exposures.

Example

Occupational Exposure (yrs)	Asbestosis		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

Therefore a person who has been exposed to asbestos for 20 years or more is:

6.396/0.39 = 16.4

OR

$$\theta = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

$$= \frac{522 \times 339}{53 \times 203}$$

$$= 16.45$$

times more likely to contract asbestosis than those who have been exposed to asbestos for less than 20 years.

Occupational Exposure (yrs)	**Asbestosis**		
	Yes	*No*	*Total*
0 – 19	203	522	725
20+	339	53	392
Total	542	575	1117

Note that if we swap the columns of the table then the odds ratio changes:

$$\theta = \frac{203 \times 53}{339 \times 522}$$
$$= 0.0608$$

Yet the overall association between the variables (regardless of direction of the association) remains unchanged.

The Log-Odds Ratio

While the strength of the association between the variables remains unchanged, its not apparent by considering the odds ratio that its only the direction of the association that's changed.

To remedy this we can consider the natural logarithm of the odds ratio, or the *log-odds ratio*: $\ln\theta$

For example:

$$\textit{Unswapped columns}: \; \ln\theta = \ln(16.45) = 2.8$$

$$\textit{Swapped columns}: \; \ln\theta = \ln(0.0608) = -2.8$$

Therefore by considering the log-odds ratio, the strength of the association remains unchanged, but the direction has been considered.

The Variance of the Log-Odds Ratio

$$\mathrm{Var}[\ln\theta] = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

In fact, the 100(1 – α)% confidence interval for the population log-odds ratio is:

$$\left(\ln\theta - Z_{\alpha/2}\mathrm{SE}[\ln\theta],\ \ln\theta + Z_{\alpha/2}\mathrm{SE}[\ln\theta]\right)$$

Example

$$\mathrm{Var}[\ln\theta] = \left(\frac{1}{522} + \frac{1}{203} + \frac{1}{53} + \frac{1}{392}\right)$$
$$= 0.0283$$

Occupational Exposure (yrs)	**Asbestosis**		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

$$\ln\theta = 2.8$$

$$\begin{pmatrix} 2.8 - 1.96 \times \sqrt{0.0283}, & \\ & 2.8 - 1.96 \times \sqrt{0.0283} \end{pmatrix} = (2.69,\ 2.90)$$

To find the variance of the odds ratio, θ, we shall consider the *Delta method*:

Suppose we have a function g(X). Then if the variance of X is known, and g(X) is twice differentiable, its variance is

$$\mathrm{Var}[g(X)] = \left(\frac{d}{dX} g(X)\right)^2 \mathrm{Var}(X)$$

So, using the Delta method when g(X) = e^X then $g(\ln\theta) = e^{\ln\theta} = \theta$ and

$$\mathrm{Var}[g(\ln\theta)] = \left(\frac{d}{d\ln\theta} e^{\ln\theta}\right)^2 \mathrm{Var}(\ln\theta)$$

simplifies to

$$\mathrm{Var}[\theta] = \left(e^{\ln\theta}\right)^2 \mathrm{Var}(\ln\theta)$$

$$= \theta^2\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

Example

Occupational Exposure (yrs)	Asbestosis		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

$$\mathrm{Var}[\theta] = \theta^2\left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

$$= (16.45)^2\left(\frac{1}{522} + \frac{1}{203} + \frac{1}{53} + \frac{1}{339}\right)$$

$$= 7.755$$

$$\theta = 16.45$$

Unlike the log-odds ratio, there is no simple procedure for finding the confidence interval of the odds ratio

Due to the widespread use of logistic regression, the odds ratio is widely used in many fields of medical and social science research.

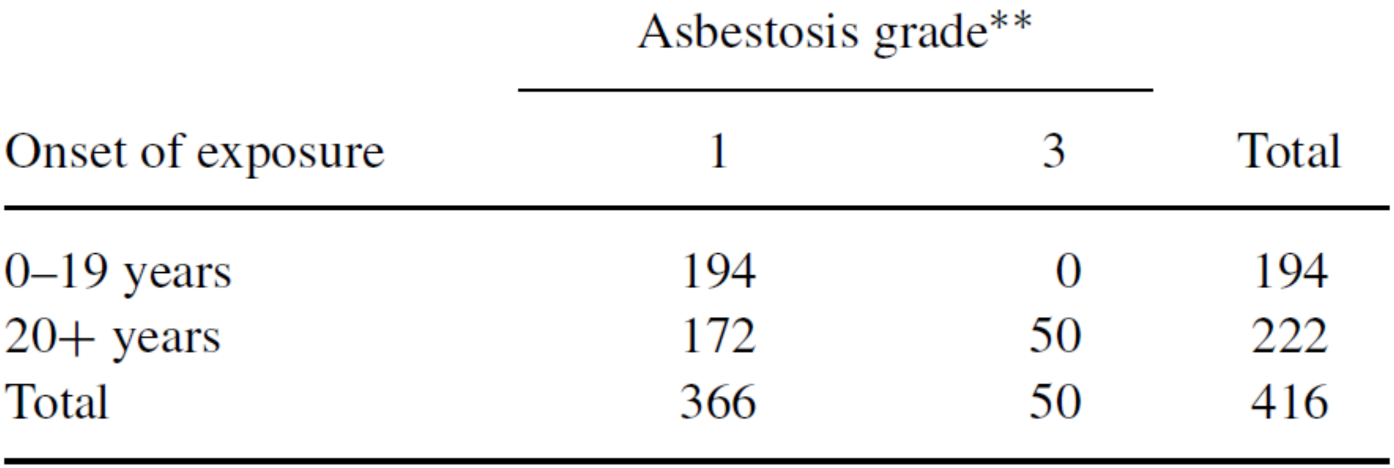
It is commonly used in survey research, in epidemiology and to express the results of some clinical trials, such as in case-control studies.

It underpins the field of meta-analysis and is now a standard approach to synthesise research findings in many disciplines, including medical and healthcare research, and climate change research and increasingly in genome-wide studies

θ and Zero Cell Frequencies

The odds ratio can only be calculated when the off diagonal elements are greater than zero.

Suppose we consider only those workers diagnosed with the least severe (grade 1) most severe case (grade 3) of asbestosis.

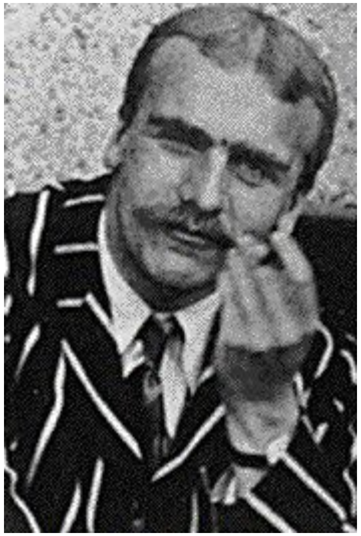


	Asbestosis grade**		
Onset of exposure	1	3	Total
0–19 years	194	0	194
20+ years	172	50	222
Total	366	50	416

$$\theta = \frac{194 \times 50}{172 \times 0} = \text{undefined}$$

We can consider alternative expressions of the odds ratio

Haldane's Odds Ratio



John B. S. Haldane
(1892–1964)

Typically, for zero cell frequencies, the analysis of categorical variables is carried out by either

- Substituting the zero cell's with a small value, usually 0.5
- Adding a small value (usually 0.5) to ALL of the cell values

When doing the second for the odds ratio leads to Haldane's (1955) odd ratio

$$\theta_H = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}$$

Breslow (1981) points out that choosing 0.5 has become popular as it best reduces the possibility of bias in small samples.

Jewell's Odds Ratio

When n is considered small, one may consider Jewell's (1984, 1986) odds ratio

$$\theta_J = \frac{n_{11} n_{22}}{(n_{12}+1)(n_{21}+1)}$$

Nicholas P. Jewell

With the wealth of advances in measuring the association between two dichotomous variables using the odds ratio, the issue now becomes which one is the most appropriate to use. For small sample sizes, the odds ratio *θ is positively biased and overestimates the true* odds ratio (Jewell, 1984, 1986). Walter and Cook (1991) conducted a comparative study of θ, θ_H, θ_J and θ_C and concluded that *θ_H is the most appropriate estimate* since it is more stable and less biased than θ. Although it was verified that, *for small samples, θ_J is appropriate*.

Yule’s Q

We shall now tie in the old established links between Pearson’s tetrachoric correlation and Cornfield’s odds ratio.

George Udny Yule
(1871 - 1951)

Recall that one approximation of the tetrachoric correlation was Yule’s Q:

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

Yule referred to this Q as the *coefficient of variation*.

Interesting, Yule (1912, pg 586) says of his notation

> *“I took the symbol from the initial letter of Quetelet”*

which is also why Pearson referred to his better approximations of his tetrachoric correlation by the letter Q.

Note that Yule's Q may be expressed in terms of the odds ratio such that

$$Q = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$
$$= \frac{(n_{11}n_{22})/(n_{12}n_{21}) - 1}{(n_{11}n_{22})/(n_{12}n_{21}) + 1}$$
$$= \frac{\theta - 1}{\theta + 1}$$

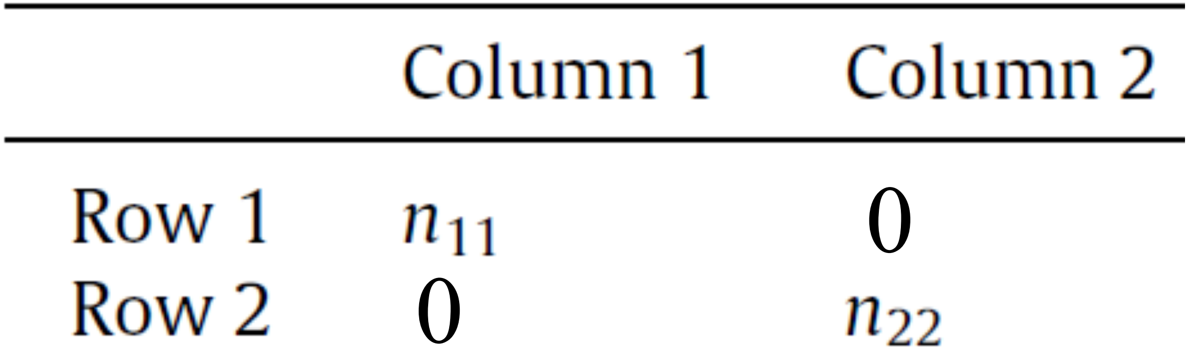
While its outside of the scope of this lecture, Yule's Q for multiple dichotomous variables was considered by Lipsitz and Fitzmaurice (1994).

The key properties of Yule’s Q are

- If there is complete independence between the row and column variables, so that $p_{ij} = p_{i\bullet}p_{\bullet j}$, for i, j = 1, 2, Q = 0
- When $n_{12} = 0$, or $n_{21} = 0$ (or both) then Q = 1. When, the off diagonal elements of the contingency table are zero there is a perfect positive association between the two dichotomous categorical variables. This property also overcomes the problem of the odds ratio which, in this case, is undefined.
- Yule’s Q ranges from –1 to +1, where a zero reflects no association between the categorical variables. Q = 0 arises when $n_{11}\, n_{11} = n_{12}\, n_{21}$ and is consistent with an odds ratio of $\theta = 1$.

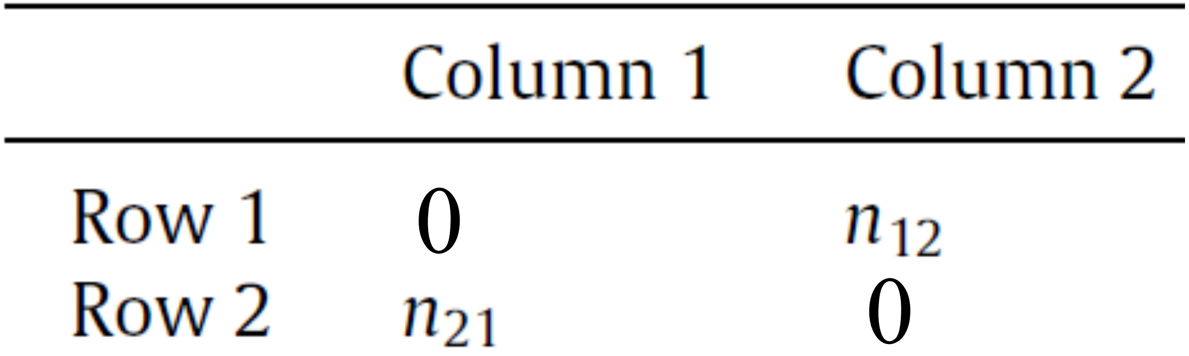
Yule's Q

- When $n_{12} = 0$, or $n_{21} = 0$ (or both), then Q = +1.



	Column 1	Column 2
Row 1	n_{11}	0
Row 2	0	n_{22}

- When $n_{11} = 0$, or $n_{22} = 0$ (or both), then Q = –1.



	Column 1	Column 2
Row 1	0	n_{12}
Row 2	n_{21}	0

Yule’s Y

In May 1912, Yule published is “coefficient of colligation”

$$Y = \frac{\sqrt{n_{11}n_{22}} - \sqrt{n_{12}n_{21}}}{\sqrt{n_{11}n_{22}} + \sqrt{n_{12}n_{21}}}$$

which is now referred to as *Yule’s Y*.

As for Yule’s preference, Yule (1912, pg 592) comments that

> *“ . . . I should be inclined to prefer [Y] to Q for any future work”*

Its link to the odds ratio is easily established:

$$Y = \frac{\sqrt{(n_{11}n_{22})/(n_{12}n_{21})} - 1}{\sqrt{(n_{11}n_{22})/(n_{12}n_{21})} + 1} = \frac{\sqrt{\theta} - 1}{\sqrt{\theta} + 1}$$

The Pearson/Yule War

Yule (1912, pg 585) concedes of his Q . . .

> *“The expression was not derived by any extraneous considerations, but was simply written down as an empirical formula fulfilling the required conditions . . .”*

Such an informal approach drew condemnation from various statisticians, especially from Yule’s mentor Karl Pearson. Despite the good working relationship the pair had in the beginning, Pearson and Yule’s work on correlation and association diverged to a point where they vehemently attacked each other’s work.

Following Yule's proposal of his Q statistic Pearson and Heron (1913) stated

> *"Naturally when one finds a method wholly inadequate one does not turn and rend an old pupil and former colleague. What Pearson did do was to test Mr Yule's Q against other similar coefficients and finding it less stable than any of them, it was dropped and has never been and never will be used in any work done under his supervision."*

Such personal and scathing remarks dominate Pearson and Heron's (1913) 157 page review of Yule's contribution.

Pearson and Heron (1913, page 160) continued their critical appraisal of Yule's contribution to statistics by saying

> *". . . if Mr Yule's views are accepted, irreparable damage will be done to the growth of modern statistical theory. Mr Yule has invented a series of statistical methods which are in no case based on a reasoned theory"*

The passage of time has now shown that both Pearson and Yule had valid points on both sides of the argument; see, for example, Agresti's (2002, page 621) excellent discussion on this history.

They also helped to change how contributions to journals are dealt with – from a forum for where personal attacks were allowed to flourish to one where the writing style is confined only the objective scientific merits of the contribution.

Pearson vs Yule

Despite Pearson's very public condemnation of Yule's work

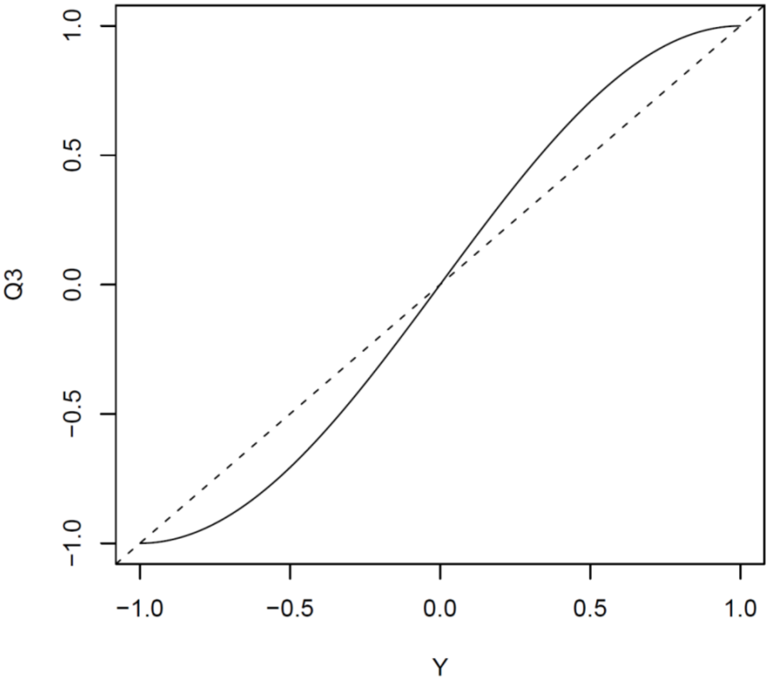
- Pearson considered a normally distributed random variable that has been dichotomised to form two a variable with two categories. Therefore, Pearson's work involved heavily theoretical work involving the normal distribution.

- Yule did not think in this way at all. Instead, his aim was to very simple, intuitive, and easily calculable **approximations** of measures of association that cross over to Pearson's work.

Yule's Y is a good approximation of Pearson's Q_3,

$$Q_3 = \sin\left(\frac{\pi}{2}Y\right)$$

especially when Y is at its extremes (close to 1 or + 1, or near zero).



Digby’s H

One will note that Yule’s Q and Yule’s Y are of the same form except for the power:

$$Q = \frac{\theta - 1}{\theta + 1} \qquad Y = \frac{\theta^{1/2} - 1}{\theta^{1/2} + 1}$$

This may suggest that some power other than 1 or 0.5 could be used as a measure of association.

Digby (1983, page 754) proposed an alternative measure of association

$$H = \frac{\theta^{3/4} - 1}{\theta^{3/4} + 1}$$

and went on to say

> *“Investigations of powers of [θ] other than 0.75 shows that any improvement over H would be minimal”*

Digby's H & Edwards J

Note that Digby's H is a special case of Yule's Y since

$$H = \frac{\sqrt{\theta^{3/2}} - 1}{\sqrt{\theta^{3/2}} + 1}$$

Therefore, as part of Digby's study, he demonstrated that H provides a compromise between Yule's Q and Y since H lies within the interval [Q, Y].

A related measure is

$$J = \frac{\theta^{\pi/4} - 1}{\theta^{\pi/4} + 1}$$

and was considered earlier by Edwards (1957). Edwards and Edwards (1984) point out that Digby's results performed well since the power 3/4 is a simple approximation of π/4.

Edwards' Criteria

Anthony WF Edwards
(1935 -)

An important feature of the odds ratio is that its magnitude remains unchanged even when the row and column categories are interchanged (Edwards, 1963). This is an important property since it ensures that the association structure remains preserved in the event that the "ordering" of row (and column) categories is reversed.

As a result, Edwards (1963) proposed that any reasonable measure of association should be expressible in terms of the odds ratio to preserve this characteristic.

Let f (θ) be a monotonically increasing function in terms of the odds ratio, θ. Bishop, Fienberg and Holland (1975, page 378) considered the function

$$g(\theta)=\frac{f(\theta)-1}{f(\theta)+1}$$

Note: Edwards is one of Britain's most distinguished geneticists and studied genetics at Cambridge as one of the last students of R. A. Fisher, Here we have considered $f(\theta) = \theta^b$

$$g(\theta) = \frac{f(\theta) - 1}{f(\theta) + 1}$$

For example:

Measure	b
Yule's Q	1
Yule's Y	1/2
Digby's H	3/4
Edwards' J	π/4

$$Q = \frac{\theta - 1}{\theta + 1} \qquad Y = \frac{\theta^{1/2} - 1}{\theta^{1/2} + 1}$$

$$H = \frac{\theta^{3/4} - 1}{\theta^{3/4} + 1} \qquad J = \frac{\theta^{\pi/4} - 1}{\theta^{\pi/4} + 1}$$

Edwards' Criteria

Occupational Exposure (yrs)	Asbestosis		
	No	*Yes*	*Total*
0 – 19	522	203	725
20+	53	339	392
Total	575	542	1117

```
> oddsratio.exe <- function (N, acc = 3) {
+ round((N[1,1]*N[2,2])/(N[1,2]*N[2,1]),
+ digits = acc)
+ }
> oddratio.exe(asbestos.dat)
> 16.447
>
> edwards.odds.exe <- function (N, b = 1, acc = 3) {
+ OR <- oddsratio.exe(N, acc)
+ round((OR^b - 1)/(OR^b + 1), digits = acc)
+ }
>
> # Yule's Q
> edwards.odds.exe(selikoff.dat, acc = 4)
[1] 0.7304
>
> # Yule's Y
> edwards.odds.exe(selikoff.dat, b = 0.5, acc = 4)
[1] 0.434
>
> # Digby's H
> edwards.odds.exe(selikoff.dat, b = 3/4, acc = 4)
[1] 0.6026
>
> # Edward's J
> edwards.odds.exe(selikoff.dat, b = pi/4, acc = 4)
[1] 0.6231
>
```

The Variance & Edwards Criteria

Bishop, Fienberg and Holland (1975, pg 378) showed that

$$\mathrm{Var}[g(\theta)] = \frac{(1-g(\theta))^4}{4}\left(\frac{d}{d\theta}f(\theta)\right)^2 \mathrm{Var}(\theta)$$

Proof

See Question 1 ii) of Assignment 1

From this, we can obtain the variance of Yule's Q, Yule's Y, Digby's H and Edwards J – these variances can be incorporated into the R code given a few slides ago.

Example

$$\mathrm{Var}[Q] = \frac{1}{4}\left(1-Q^2\right)^2\left(\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right)$$

$$\mathrm{Var}[Y] = \frac{1}{16}\left(1-Y^2\right)^2\left(\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}\right)$$

Common Odds Ratio

When we have MULTIPLE (G > 1) 2x2 contingency tables, we could assess the association structure between the two dichotomous variables by

- Finding the odds ratio of each of the G 2x2 tables
- Determine the odds ratio that is “common” to all of the G tables

Here we will look at ways in which we can calculate the “common” odds ratio for multiple 2x2 tables. There are two popular measures we can consider.

The

- Mantel-Haenszel (MH) odds ratio
- Woolf’s odds ratio

We will look directly at the MH odds ratio. I shall leave you to find out about the Woolf estimator in Assignment 1.

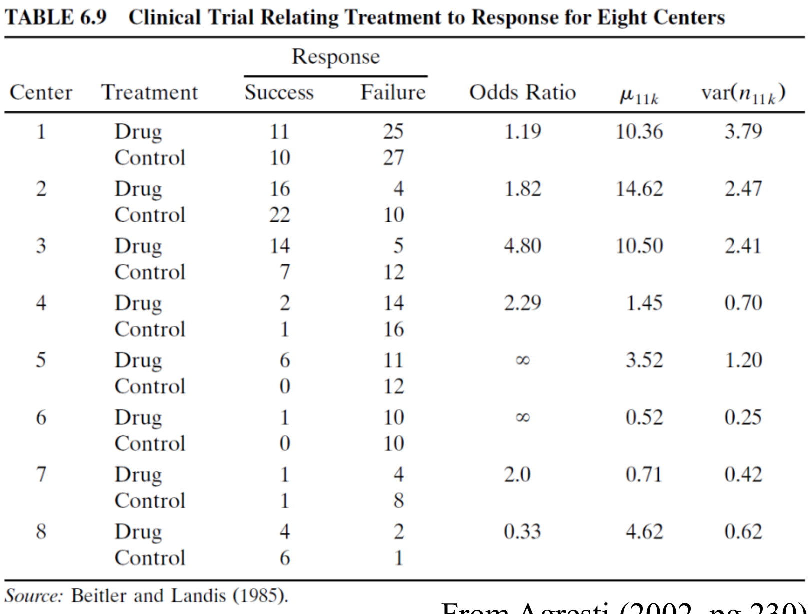


TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers

		Response				
Center	Treatment	Success	Failure	Odds Ratio	μ_{11k}	var(n_{11k})
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

From Agresti (2002, pg 230)

Common Odds Ratio

The Mantel-Haenszel Odds Ratio

Mantel and Haenszel (1959) proposed a testing procedure for

H_0: conditional independence of the two dichotomous variable of 2x2xG tables

We will talk about the test soon, but we first calculate their “common” odds ratio.

For the (i, j)th cell of the g table, n_{ijg}, given the row and column marginal totals, the probability of observing this count can be determined using the hypergeometric distribution (see STAT2010 lecture notes). That is, for the (1, 1)th cell

$$P\left(X = n_{11g}\right) = \frac{\binom{n_{1\cdot g}}{n_{11g}}\binom{n_{2\cdot g}}{n_{\cdot 1g}-n_{11g}}}{\binom{n_g}{n_{\cdot 1g}}},$$

	Y1	Y2	Total
X1	n_{11g}	n_{12g}	$n_{1\cdot g}$
X2	n_{21g}	n_{22g}	$n_{2\cdot g}$
Total	$n_{\cdot 1g}$	$n_{\cdot 2g}$	n_g

Notation for the g’th 2x2 table

The Mantel-Haenszel odds ratio is just the weighted mean of each of the G odds ratio’s:

$$\hat{\theta}_{MH} = \frac{\sum_{g=1}^{G}\left(n_{11g}n_{22g}/n_g\right)}{\sum_{g=1}^{G}\left(n_{12g}n_{21g}/n_g\right)} = \frac{\sum_{g=1}^{G} R_g}{\sum_{g=1}^{G} S_g} = \frac{R}{S},$$

Common Odds Ratio

Robins, Breslow and Greenland (1986) showed various characteristics of the Mantel-Haenszel odds ratio estimator. In particular, they showed that the variance of the log Mantel-Haenszel estimator is

$$
\begin{aligned}
(\sigma_{MH})^2 &= \mathrm{Var}\left(\ln\hat{\theta}_{\mathrm{MH}}\right) \\
&= \frac{1}{2R^2}\sum_{g=1}^{G}\frac{\left(n_{11g}+n_{22g}\right)R_g}{n_g}+\frac{1}{2S^2}\sum_{g=1}^{G}\frac{\left(n_{12g}+n_{21g}\right)S_g}{n_g} \\
&\qquad +\frac{1}{2RS}\sum_{g=1}^{G}\frac{\left(n_{11g}+n_{22g}\right)R_g+\left(n_{12g}+n_{21g}\right)S_g}{n_g}
\end{aligned}
$$

They also showed that $\ln\hat{\theta}_{\mathrm{MH}}$ is asymptotically normally distributed with mean $\ln\theta$ (where θ is the common population odds ratio) and variance $(\sigma_{MH})^2$

Common Odds Ratio

Example

Suppose we consider the example given by Agresti (2002, pg 230)

$$\hat{\theta}_{MH} = \frac{\sum_{g=1}^{G}(n_{11g}n_{22g}/n_g)}{\sum_{g=1}^{G}(n_{12g}n_{21g}/n_g)}$$
$$= \frac{(11 \times 27)/73 + (16 \times 10)/52 + \cdots + (4 \times 1)/13}{(10 \times 25)/73 + (22 \times 4)/52 + \cdots + (6 \times 2)/13}$$
$$= 2.13$$

So $\quad \ln\hat{\theta}_{MH} = 0.758.$

Also, $\quad (\sigma_{MH})^2 = 0.303$

So, the 95% confidence interval for the common (Mantel-Haenszel) **log-odds** ratio is $0.758 \pm 1.96 \times 0.303 = (0.164, 1.352)$.

Therefore the 95% confidence interval for the Mantel-Haenszel odds ratio is

$$\exp(0.164, 1.352) = (1.178, 3.865)$$

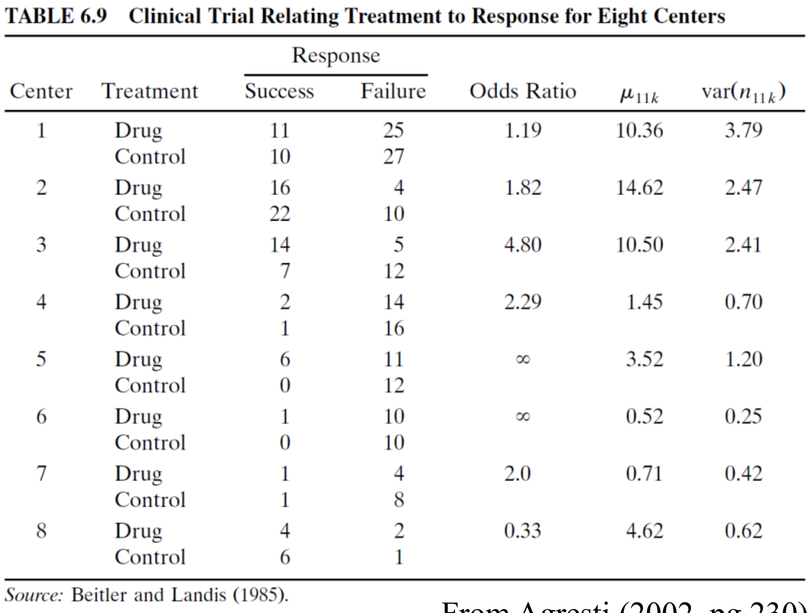


TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers

Center	Treatment	Response Success	 Failure	Odds Ratio	μ_{11k}	var(n_{11k})
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

From Agresti (2002, pg 230)

The Cochran-Mantel-Haenszel Test

Mantel and Haenszel (1959) originally proposed a testing procedure for testing homogenous independence across the G 2x2 tables

$$H_0\colon \theta_1 = \theta_2 = \cdots = \theta_G = 1$$

We adopt the same notation we have used so far. For the (i, j)th cell of the g table, n_{ijg}, given the row and column marginal totals, the probability of observing this count can be determined using the hypergeometric distribution (see STAT2010 lecture notes). That is, for the (1, 1)th cell

$$P(X = n_{11g}) = \frac{\binom{n_{1\cdot g}}{n_{11g}}\binom{n_{2\cdot g}}{n_{\cdot 1g}-n_{11g}}}{\binom{n_g}{n_{\cdot 1g}}},$$

	Y1	Y2	Total
X1	n_{11g}	n_{12g}	$n_{1\cdot g}$
X2	n_{21g}	n_{22g}	$n_{2\cdot g}$
Total	$n_{\cdot 1g}$	$n_{\cdot 2g}$	n_g

Notation for the g’th 2x2 table

So, under independence, and from the hypergeometric distribution

$$E(n_{11g}) = \frac{n_{1\cdot g}n_{\cdot 1g}}{n_g} \qquad Var(n_{11g}) = \frac{n_{1\cdot g}n_{2\cdot g}n_{\cdot 1g}n_{\cdot 2g}}{(n_g)^2(n_g - 1)}$$

Note that independence here assumes that, at each 2x2 table, the odds ratio is 1.

Therefore, the statistic

$$\mathrm{CMH} = \frac{\sum_{g=1}^{G}\left(n_{11g} - E\left(n_{11g}\right)\right)}{\sum_{g=1}^{G} \mathrm{Var}\left(n_{11g}\right)}$$

is commonly referred to as the Cochran-Mantel-Haenszel statistic and, for large samples, has a chi-squared distribution with 1 degree of freedom.

Where does Cochran fit into this??

Cochran (1954) proposed a similar statistic but considered that the two rows (which are treated as being independent) are binomial, rather than hypergeometric. So he considered the variance

$$\mathrm{Var}\left(n_{11g}\right) = \frac{n_{1\cdot g} n_{2\cdot g} n_{\cdot 1g} n_{\cdot 2g}}{\left(n_{g}\right)^{3}}$$

Which is virtually the same as the Mantel-Haenzsel variance for large sample sizes.

The Breslow-Day Test

Breslow & Day (1985), Tarone (1985)

Since null hypothesis of the CMH test assumes independence at each table, this restriction can be relaxed by considering the Breslow-Day test

$$H_0: \theta_1 = \theta_2 = \cdots = \theta_G = \theta$$

Where θ is the common population odds ratio. The test often uses the Mantel-Haenszel odds ratio as an estimator of θ but any reasonably chosen value may be used.

$$\mathrm{BD} = \sum_{g=1}^{G} \frac{\left(n_{11g} - E\left(n_{11g}|\hat{\theta}\right)\right)^2}{\mathrm{Var}\left(n_{11g}\right)}$$

	Y1	Y2	Total
X1	n_{11g}	n_{12g}	$n_{1\cdot g}$
X2	n_{21g}	n_{22g}	$n_{2\cdot g}$
Total	$n_{\cdot 1g}$	$n_{\cdot 2g}$	n_g

Notation for the g'th 2x2 table

Here

$$\mathrm{Var}\left(n_{11g}\right) = \left[\frac{1}{x_{11g}} + \frac{1}{x_{12g}} + \frac{1}{x_{21g}} + \frac{1}{x_{22g}}\right]^{-1}$$

where $x_{11g} = E\left(n_{11g}|\hat{\theta}\right)$.

We know what the expected value of n_{11g} is under independence.

So, how to find the expected value of n_{11g} for a specific odds ratio

Common Odds Ratio

The expectation of n_{ijg} given an estimate of the common odds ratio, denoted by $x_{11g} = E(n_{11g}|\hat{\theta})$, is found by solving

$$\frac{x_{11g}(n - n_{1\cdot g} - n_{\cdot 1g} + x_{11g})}{(n_{1\cdot g} - x_{11g})(n_{\cdot 1g} - x_{11g})} = \hat{\theta}$$

Think about why?

The solution to x is therefore

Think about why?

$$x_{11g} = E(n_{11g}|\hat{\theta})$$

$$= \frac{\left(n_{2\cdot g} - n_{\cdot 1g} + \hat{\theta}(n_{1\cdot g} - n_{\cdot 1g})\right) \pm \sqrt{\left(n_{2\cdot g} - n_{\cdot 1g} + \hat{\theta}(n_{1\cdot g} - n_{\cdot 1g})\right)^2 - 4\hat{\theta}(\hat{\theta} - 1)n_{1\cdot g}n_{\cdot 1g}}}{2(\hat{\theta} - 1)}$$

Once this is found, the expected value of the remaining three cell frequencies can be found by subtraction to yield the variance term. These are then substituted into BD to find the Breslow-Day statistic.

We could also consider the non-central hypergeometric distribution to find x_{11g} (we wont though)

The Breslow-Day statistic

- has approximately chi-squared distribution with df = G − 1, given large sample size, and under H_0
- it does not work well for small sample size, while CMH works fine

Check http://www.math.montana.edu/~jimrc/classes/stat524/Rcode/breslowday.test.r for an R function that performs these calculations and the Breslow-Day test

Next Week

- Next week (Week 4) we shall look at other measures of association for categorical variables summarised in the form of an IxJ contingency table.
- In doing so we shall also explore the use of measures and issues including

  - Power divergence statistic

  - Kruskal-tau index & the C statistic

- *Week 5* – Scoring methods for categorical variables (reciprocal averaging, eigen-decomposition, SVD)
- *Week 6* – Simple correspondence analysis (graphical representation of the association between two categorical variables)

References

- Agresti A 2002 *Categorical Data Analysis*. Wiley.
- Beh EJ and Smith DR 2011, Real world occupational epidemiology, Part 1: Odds ratios, relative risk and asbestos, *Archives of Environmental & Occupational Health*, 66,119 – 123.
- Bishop YMM, Fienberg SE and Holland PW 1975 *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.
- Breslow NE and Day NE 1980 *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies*. IARC, Lyon
- Brown MB 1977 Algorithm AS 116: The tetrachoric correlation and its asymptotic standard error. *Applied Statistics* 26, 343 – 351.
- Brown MB and Benedetti JK 1977 On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika* 42, 347 – 355.
- Cochran WG 1954 Some methods of strengthening the common χ2 tests. *Biometrics*, 10, 417 – 151.
- Cornfield J 1951 A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* 11, 1269 – 1275.
- Digby PGN 1983 Approximating the tetrachoric correlation coefficient. *Biometrics* 39, 753 – 757
- Edwards JH 1957 A note on the practical interpretation of 2 *x* 2 tables. *British Journal of Preventative Social Medicine* 11, 73 – 78.
- Edwards AWF 1963 The measure of association in a 2£2 table. *Journal of the Royal Statistical Society, Series A* 126, 109 – 114.
- Fleiss JL, Levin B and Paik MC 2003 *Statistical Methods for Rates and Proportions*, 3rd edn. Wiley.

References

- Fleming JS 2005 TETCORR: A computer program to compute smoothed tetrachoric correlation matrices. *Behavior Research Methods* 37, 59 – 64
- Froemel ECA 1971 A comparison of computer routines for the calculation of the tetrachoric correlation coefficient. *Psychometrika* 36, 165 – 174.
- Haldane JBS 1955 The estimation and significance of the logarithm of a ratio frequencies. *Annals of Human Genetics 20*, 309 – 311
- Jewell NP 1984 Small sample bias of point estimators of the odds ratio from matched sets. *Biometrics* 40, 421 – 435.
- Jewell NP 1986 On the bias of commonly used measures of association for 2 *x* 2 tables. *Biometrics* 42, 351 – 358.
- Lipsitz SR and Fitzmaurice G 1994 An extension of Yule's Q to multivariate binary data. *Biometrics* 50, 847 – 852.
- Mantel N and Greenhouse SW 1968 What is the continuity correction? *The American Statistician* 22 (5), 27 – 30.
- Mantel N and Haenszel W 1959 Statistical aspects of the analysis of data from restrospective studies of cancer, *Journal of the National Cancer Institute*, 22, 719 – 748.
- Martinson EO and Hamdan MA 1975 Algorithm AS87: Calculation of the polychoric estimate of correlation in contingency tables. *Applied Statistics* 24, 272 – 278
- Pearson K 1900 On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* 1, 157 – 175.
- Pearson K and Heron D 1913 On theories of association. *Biometrika* 9, 159 – 315.
- Robins J, Breslow N and Greenland S 1986 Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, 42, 311 – 323.
- Stuart A, Ord K and Arnold S 2002 *Kendall's Advanced Theory of Statistics; Classical Inference & the Linear Model*, 6th edn. Wiley.

References

- Tarone RE 1985 On heterogeneity tests based on efficient scores. *Biometrika*, 72, 91 – 95.
- Walter SD and Cook RJ 1991 A comparison of several point estimators of the odds ratio in a single 2 *x* 2 contingency tables. *Biometrics* 47, 795 – 811.
- Yates F 1934 Contingency tables involving small numbers and the test. *Journal of the Royal Statistical Society Supplement* 1, 217 – 235.
- Yule GU 1900 On the Association of Attributes in Statistics. *Philosophical Transactions of the Royal Society of London, Series A* 194, 257 – 319.
- Yule GU 1912 On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society* 75, 579 – 652