Two-Way Contingency Tables

Joint, Marginal and Conditional Distributions

Suppose X and Y are two categorical response variables, with X having I levels and Y having J levels and that we classify each item in a population using both variables 9i.e. the data is said to be cross-classified.).

Now consider a randomly chosen item from this population. The responses (X, Y) corresponding to this item have a joint probability distribution. We lLet π_{ij} denote the probability that X assumes its i^{th} level and Y assumes its j^{th} level.

Consider the following I x J table:

		Y						
		1	2	$\cdots$	j	...	J	Total
X	1	π_{11}	π_{12}	...	π_{1j}	...	π_{1J}	π_{1+}
	2	π_{21}	π_{22}	...	π_{2j}	...	π_{2J}	π_{2+}
	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	i	π_{i1}	π_{i2}	...	π_{ij}	...	π_{iJ}	π_{i+}
	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	I	π_{I1}	π_{I2}	...	π_{Ij}	...	π_{IJ}	π_{I+}
	Total	π_{+1}	π_{+2}	...	π_{+j}	...	π_{+J}	$\pi_{++} = 1$

The probability distribution $\{\pi_{ij}\}$ is the **joint distribution of** X **and** Y and defines the (bivariate) relationship between these two variables.

The **marginal distributions of** X **and** Y are respectively the row and column totals, obtained by summing the appropriate joint probabilities. These are denoted by $\{\pi_{i+}\}$ for X and $\{\pi_{+j}\}$ for Y. The marginal

distributions represent ***single-variable*** information and **do not refer to association links** between the two variables.

Generally $\{\pi_{ij}\}$, $\{\pi_{i+}\}$, and $\{\pi_{+j}\}$ are unknown but they can be estimated by sampling.

Example: Consider a sample of 1783 U.S.military veterans cross-classified by sleep problems. This yields a **2 x 2 contingency table**.

		Sleep Problems		
		Yes	No	Total
Service in military	Yes	$n_{11} = 173$	$n_{12} = 599$	$n_{1+} = 772$
	No	$n_{21} = 160$	$n_{22} = 851$	$n_{2+} = 1011$
	Total	$n_{+1} = 333$	$n_{+2} = 1450$	$n_{++} = 1783$

Note: Here the *overall sample size is fixed* ***but*** *row and column totals are not fixed*. Thus this study corresponds to a multinomial sample with 4 outcomes.

The maximum likelihood estimates (M.L.E.s) of $\{\pi_{ij}\}$, $\{\pi_{i+}\}$and $\{\pi_{+j}\}$ are $\{p_{ij}\}$, $\{p_{i+}\}$and $\{p_{+j}\}$ respectively and are given below:

		Sleep Problems: Yes	No	Total
Service in military	Yes	$p_{11} = 0.097$	$p_{12} = 0.336$	$p_{1+} = 0.433$
	No	$p_{21} = 0.090$	$p_{22} = 0.447$	$p_{2+} = 0.567$
	Total	$p_{+1} = 0.187$	$p_{+2} = 0.813$	$p_{++} = 1.000$

In some cases one variable can be thought of as a *response* variable and the other as an *explanatory* variable. (In this study, we might treat sleep problems as a response variable and service in the military as an explanatory variable.) For such cases, it is useful to construct a separate probability distribution for Y at each level of X. Given that an item is classified in row i of X, we use $\pi_{j|i}$ to denote the probability of classification in column j of Y. This yields the following table:

		Y: 1	2	...	j	...	J	Total
X	1	$\pi_{1\|1}$	$\pi_{2\|1}$	...	$\pi_{j\|1}$	...	$\pi_{J\|1}$	1
	2	$\pi_{1\|2}$	$\pi_{2\|2}$	...	$\pi_{j\|2}$	...	$\pi_{J\|2}$	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
	i	$\pi_{1\|i}$	$\pi_{2\|i}$	...	$\pi_{j\|i}$	...	$\pi_{J\|i}$	1
	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$
	I	$\pi_{1\|I}$	$\pi_{2\|I}$	...	$\pi_{j\|I}$	...	$\pi_{J\|I}$	1

The probabilities $\{\pi_{1|i}, \pi_{2|i}, ..., \pi_{J|i}\}$ represent the ***conditional*** *distribution of Y at the i^{th} level of X*. The conditional distribution of Y given $X = i$ is related to the joint distribution of $\{X, Y\}$ by

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \text{ for all } i \text{ and } j.$$

Usually these conditional probability distributions are also unknown and can be estimated by sampling.

For our example, we estimate the conditional probability distribution for sleep problems $\{\pi_{1|i}, \pi_{2|i}\}$ at the i^{th} level of military service using $\{p_{1|i}, p_{2|i}\}$. These conditional probabilities are shown below:

		Sleep Problems: Yes	No	Total
Service in military	Yes	$p_{1\|1} = 0.224$	$p_{2\|1} = 0.776$	1
	No	$p_{1\|2} = 0.158$	$p_{2\|2} = 0.842$	1

Independence

When both variables are *response* variables, we can describe their association using:
- their *joint* distribution,
- the *conditional distribution of Y given X*
- the *conditional distribution of X given Y*.

The variables X and Y are said to be **statistically independent** if

$$\pi_{ij} = \pi_{i+}\pi_{+j} \text{ for } i = 1,...,I \text{ and } j = 1,...,J$$

Thus, when X and Y are independent, we have that for each $j = 1,...,J$

$$\pi_{j|i} = P(Y = j \mid X = i) = \frac{P(Y = j, X = i)}{P(X = i)}$$

$$= \frac{\pi_{ij}}{\pi_{i+}} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j}$$

for $i = 1,...,I$.

However when Y is a *response* and X is an *explanatory* variable, the condition

$$\pi_{j|1} = \pi_{j|2} = ... = \pi_{j|I}$$

for all j is a more natural definition of independence.

Note: In some tables where Y is a response variable and X is an explanatory variable, X is **fixed** rather than random. In such cases the idea of a joint distribution for X and Y is no longer meaningful. However, for a fixed level of X, Y still has a probability distribution. We would therefore consider the conditional distribution of Y and different fixed levels of X.

Test for Homogeneity: Prospective Study

Example:

The Physicians' Health Study was a 5 year study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. In this study, 22,071 physicians were randomly assigned either to a group that was to take one aspirin tablet every other day or to a group that was to take a placebo every other day. Of the 22,071 physicians, 11,034 were assigned to receive the placebo and 11,037 were assigned to receive aspirin. The study was blind - i..e. the physicians did not know which type of pill they were assigned to take.
(NOTE: This study is a **clinical trial**, since the researchers assign the physicians to the placebo and aspirin groups. Another type of prospective study is a **cohort study**, where the researchers do **not** assign individuals to groups. e.g. to study the effect of smoking on MI, a researcher might select a sample of smokers independently of a sample of nonsmokers, but the researcher does not *assign* individuals to the smoking and nonsmoking groups.)

Of the 11,034 physicians taking the placebo, 189 suffered myocardial infarcation (MI) over the course of the study while of the 11,037 taking aspirin, 104 suffered MI. The results are summarized in the following 2 x 2 **contingency table**:

		MI		
		Yes	No	Total
Group	Placebo	$n_{11} = 189$	$n_{12} = 10845$	$n_{1+} = 11034$
	Aspirin	$n_{21} = 104$	$n_{22} = 10933$	$n_{2+} = 11037$
	Total	$n_{+1} = 293$	$n_{+2} = 21778$	$n_{++} = 22071$

Note that the row (group) totals are fixed by the study.

Research Question:
Is the proporton of physicians taking a placebo who suffer MI the same as the proportion of physicians taking aspirin who suffer MI?
This is an example of a *prospective* study. (*Note*: In a prospective study, the *row totals are fixed*.)

Let $\pi_{1|1}$ = probability of suffering MI (i.e. $Y = 1$) given that the physician takes the placebo (i.e. $X = 1$)
$\pi_{2|1}$ = probability of not suffering MI (i.e. $Y = 2$) given that the physician takes the placebo (i.e. $X = 1$)
$\pi_{1|2}$ = probabiility of suffering MI (i.e. $Y = 1$) given that the physician takes aspirin (i.e. $X = 2$)
$\pi_{2|2}$ = probability of not suffering MI (i.e. $Y = 2$) given that the pysician takes aspirin (i.e. $X = 2$)

		MI. Y		
		Yes	No	Total
Group, X	Placebo	$\pi_{1\|1}$	$\pi_{2\|1}$	1
	Aspirin	$\pi_{1\|2}$	$\pi_{2\|2}$	1

The research question translates to wanting to test

$$H_0 : \pi_{1|1} = \pi_{1|2} = \pi$$

First we find MLEs for the $\pi_{j|i}$. This will allow us to determine estimated expected frequencies $\widehat{m}_{ij}$ and compare them with what we have observed. Pearson's Chi-square test statistic can then be used here.

The *likelihood* of the data is the probability of observing the sample result we have obtained and can be written as

$$\binom{n_{1+}}{n_{11}} \pi_{1|1}^{n_{11}} (1-\pi_{1|1})^{n_{1+}-n_{11}} \binom{n_{2+}}{n_{21}} \pi_{1|2}^{n_{21}} (1-\pi_{1|2})^{n_{2+}-n_{21}}$$

so the *kernel of the likelihood* is .

$$\pi_{1|1}^{n_{11}} (1-\pi_{1|1})^{n_{1+}-n_{11}} \pi_{1|2}^{n_{21}} (1-\pi_{1|2})^{n_{2+}-n_{21}}$$

Under H_0 the kernel can be rewritten and becomes

$$\begin{aligned}\pi^{n_{11}}(1-\pi)^{n_{1+}-n_{11}}\pi^{n_{21}}(1-\pi)^{n_{2+}-n_{21}} &= \pi^{n_{11}+n_{21}}(1-\pi)^{(n_{1+}+n_{2+})-(n_{11}+n_{21})} \\ &= \pi^{n_{+1}}(1-\pi)^{n-n_{+1}}\end{aligned}$$

The *log likelihood of the kernel* is

$$L = n_{1+}\log(\pi) + (n - n_{+1})\log(1-\pi)$$

and maximizing this we obtain

$$\begin{aligned}\frac{\partial L}{\partial \pi} &= \frac{n_{+1}}{\pi} - \frac{(n-n_{+1})}{1-\pi} = 0 \\ \rightarrow \quad \hat{\pi} &= \frac{n_{+1}}{n} = p_{+1}\end{aligned}$$

Thus, ***under*** H_0, $\pi_{1|1}$ and $\pi_{1|2}$ are estimated by

$$\hat{\pi} = \frac{n_{+1}}{n} = p_{+1.}$$

Hence $\pi_{2|1}(= 1-\pi_{1|1})$ and $\pi_{2|2}\ (= 1-\pi_{1|1})$ are estimated by

$$1-\hat{\pi} = \frac{n_{+2}}{n} = p_{+2}$$

Using these results, the estimated frequencies under the assumption of H_0 are

$$\hat{m}_{ij} = n_{i+}p_{+j} = n_{i+}\left(\frac{n_{+j}}{n}\right) = n_{i+}n_{+j}/n$$

Thus for our data we obtain:

$$\begin{aligned}\hat{m}_{11} &= 11034(293)/22071 = 146.48 \\ \hat{m}_{12} &= 11034(21778)/22071 = 10887.52 \\ \hat{m}_{21} &= 11037(293)/22071 = 146.52 \\ \hat{m}_{22} &= 11037(21778)/22071 = 10890.48\end{aligned}$$

Pearson's X^2 can be used to test the null hypothesis here. Recall that for large samples, $X^2 \sim \chi^2$.

Here we have

$$\begin{aligned}X^2 &= \sum\sum \frac{(n_{ij}-\hat{m}_{ij})^2}{\hat{m}_{ij}} = \frac{(189-146.48)^2}{146.48} + \frac{(10845-10887.52)^2}{10887.52} \\ &\quad + \frac{(104-146.52)^2}{146.52} + \frac{(10933-10890.48)^2}{10890.48} \\ &= 25.01\end{aligned}$$

with $df = 2-1 = 1$. The p-value is approximately 0, so there is strong evidence against H_0.

A *likelihood ratio Chi-square test* could **also** be used here.

First we maximize the likelihood under H_0; then we maximize the likelihood under $H_0 \cup H_A$.

The likelihood ratio test is based on Λ which is the ratio of the max. likeliood under H_0 to the max. likelihood.under $H_0 \cup H_A$.
For the test for homogeneity, recall that the kernel of the likelihood is

$$\pi_{1|1}^{n_{11}}(1-\pi_{1|1})^{n_{1+}-n_{11}}\pi_{1|2}^{n_{21}}(1-\pi_{1|2})^{n_{2+}-n_{21}}$$

When H_0 is assumed to be true, the kernel simplifies to

$$\pi^{n_{+1}}(1-\pi)^{n-n_{+1}}$$

and the log likelihood of this kernel is maximized at

$$\hat{\pi} = \frac{n_{+1}}{n} = p_{+1.}$$

v
Consider now the kernel *in the general context* (i.e. under $H_0 \cup H_A$). The log likelihood of this kernel is

$$L = n_{11}\log(\pi_{1|1}) + (n_{1+}-n_{11})\log(1-\pi_{1|1}) + n_{21}\log(\pi_{1|2}) + (n_{2+}-n_{21})\log(1-\pi_{1|2}).$$

We require estimates for $\pi_{1|1}$ and $\pi_{1|2}$. Now maximizing, we get

$$\frac{\partial L}{\partial \pi_{1|1}} = \frac{n_{11}}{\pi_{1|1}} - \frac{(n_{1+}-n_{11})}{1-\pi_{1|1}} = 0$$

$$\rightarrow \widehat{\pi_{1|1}} = \frac{n_{11}}{n_{1+}} = p_{1|1}$$

and similarly,

$$\frac{\partial L}{\partial \pi_{1|2}} = 0$$

$$\rightarrow \widehat{\pi_{1|2}} = \frac{n_{21}}{n_{2+}} = p_{1|2}$$

and this gives the likelihood ratio test statistic as

$$\Lambda = \frac{(\frac{n_{+1}}{n})^{n_{+1}}(1-\frac{n_{+1}}{n})^{n-n_{+1}}}{(\frac{n_{11}}{n_{1+}})^{n_{11}}(1-\frac{n_{11}}{n_{1+}})^{n_{1+}-n_{11}}(\frac{n_{21}}{n_{2+}})^{n_{21}}(1-\frac{n_{21}}{n_{2+}})^{n_{2+}-n_{21}}}$$

$$= \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(n_{i+}\ n_{+j})^{n_{ij}}}{n^n \prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}^{n_{ij}}}$$

$$= \prod_{i=1}^{I}\prod_{j=1}^{J}\left(\frac{\widehat{m}_{ij}}{n_{ij}}\right)^{n_{ij}} \text{. since } \widehat{m}_{ij} = \frac{n_{i+}n_{+j}}{n}$$

Wilks' statistics is $G^2 = -2\log\Lambda$..

For this example,

$$\begin{aligned} G^2 &= -2\log\Lambda \\ &= 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(n_{ij}/\widehat{m}_{ij}) \quad \text{where} \quad \widehat{m}_{ij} = \frac{n_{i+}n_{+j}}{n} \end{aligned}$$

and the $df = 2 - 1 = 1$ (which is the same as for Pearson's Chi-square test).

For our example,

$$\begin{aligned} G^2 &= 2\{189\log(\frac{189}{146.48}) + 10845\log(\frac{10845}{10877.52}) + 104\log(\frac{104}{146.52}) + 10933\log(\frac{10933}{10890.48})\} \\ &= 25.37 \end{aligned}$$

with p –value of approximately 0. We again would conclude that there is strong evidence against H_0

Now let's try to understand the nature of this differen ce in proportions of physicians taking aspirin who suffer MI and those physicians taking a placebo who suffer MI. To do this, we examine

- ***confidence intervals***,
- ***relative risk***, and
- ***odds ratios***.

Large Sample Confidence Interval for $\pi_{1|1} - \pi_{1|2}$:

Recall that the MLEs of $\pi_{1|1}$ and $\pi_{1|2}$ were

$$\widehat{\pi_{1|1}} = \frac{n_{11}}{n_{1+}} = p_{1|1}$$

$$\text{and } \widehat{\pi_{1|2}} = p_{1|2}$$

where n_{1+} and n_{2+} are fixed.

Also, n_{11} and n_{21} are **independent binomial random variables** with means and variances

$$E(n_{11}) = n_{1+}\pi_{1|1}$$

and

$$E(n_{21}) = n_{2+}\pi_{1|2}$$

$$Var(n_{11}) = n_{1+}\pi_{1|1}(1-\pi_{1|1})$$

and

$$Var(n_{21}) = n_{2+}\pi_{1|2}(1-\pi_{1|2})$$

Therefore $p_{1|1}$ and $p_{1|2}$ are also independent with means and variances

$$E(p_{1|1}) = E(\frac{n_{11}}{n_{1+}}) = \pi_{1|1}$$

and

$$E(p_{1|2}) = \pi_{1|2}$$

$$Var(p_{1|1}) = \pi_{1|1}(1-\pi_{1|1})/n_{1+}$$

and

$$Var(p_{1|2}) = \pi_{1|2}(1-\pi_{1|2})/n_{2+}$$

To estimate $\pi_{1|1} - \pi_{1|2}$ we can use $p_{1|1} - p_{1|2}$ as the point estimator, where

$$E(p_{1|1} - p_{1|2}) = \pi_{1|1} - \pi_{1|2}$$

$$Var(p_{1|1} - p_{1|2}) = (\pi_{1|1}(1-\pi_{1|1})/n_{1+}) + (\pi_{1|2}(1-\pi_{1|2})/n_{2+})$$

For large samples, we may use the fact that $p_{1|1}$ and $p_{1|2}$ will be *approximately* normally distributed.

Therefore a $100(1-\alpha)\%$ confidence interval for $\pi_{1|1} - \pi_{1|2}$ can be given by

$$(p_{1|1} - p_{1|2}) \pm z_{a/2}\sqrt{(p_{1|1}(1-p_{1|1})/n_{1+}) + (p_{1|2}(1-p_{1|2})/n_{2+})}$$

For our example, we may wish to obtain a a 95% confidence interval for $\pi_{1|1} - \pi_{1|2}$.,We use the fact that

$$p_{1|1} = \frac{n_{11}}{n_{1+}} = 189/11034 = 0.0171$$

and

$$p_{1|2} = \frac{n_{21}}{n_{2+}} = 104/11037 = 0.0094$$

so

$$p_{1|1} - p_{1|2} = 0.0171 - 0.0094 = 0.0077$$

and

$$\begin{aligned} &\sqrt{(p_{1|1}(1-p_{1|1})/n_{1+}) + (p_{1|2}(1-p_{1|2})/n_{2+})} \\ &= \sqrt{0.0171(1-0.0171)/11034 + 0.0094(1-0.0094)/11037} \\ &= 0.0015 \end{aligned}$$

So a 95% confidence interval for $\pi_{1|1} - \pi_{1|2}$ is given by

$$0.0077 \pm 1.96(0.0015)$$

or

$$(0.0048, 0.0106)$$

This interval does not contain 0. In fact it contains values that are > 0 , thereby indicating that aspirin appears to diminish the risk of MI.

Relative Risk

A difference between two proportions may have greater importance when both proportions are near 0 or 1 than when they are near 0.5. So, instead of studying the effect of aspirin on MI by considering the difference $\pi_{1|1} - \pi_{1|2}$, we could look at the *relative risk*, which is the ratio of the "success" probabilities (i.e. $Y = 1$) for the 2 groups. Thus we have that

$$(Population)RelativeRisk = \frac{P(Y = 1 \mid X = 1)}{P(Y = 1 \mid X = 2)} = \frac{\pi_{1|1}}{\pi_{1|2}}$$

If our H_0 is true, then this would translate as $\pi_{1|1} = \pi_{1|2}$ (i.e. the response is not affected by the group) or alternatively $\frac{\pi_{1|1}}{\pi_{1|2}} = 1$.

We would use the *sample relative risk* $\frac{p_{1|1}}{p_{1|2}}$ to estimate the population relative risk. For our example, the sample relative risk $\frac{p_{1|1}}{p_{1|2}} = \frac{0.0171}{0.0094} = 1.82$. This implies that the sample proportion of MI cases was 82% higher for the group taking the placebo than for the group taking aspirin. In other words, there is substantial evidence that taking aspirin is associated with a lower risk of having MI.

Obtaining a $100(1-\alpha)\%$ *confidence interval for the (population)relative risk* $\frac{\pi_{1|1}}{\pi_{1|2}}$

We want to base this confidence interval on the best estimator of $\frac{\pi_{1|1}}{\pi_{1|2}}$ which is $\frac{p_{1|1}}{p_{1|2}}$:

The problem here is that the distribution of $\frac{p_{1|1}}{p_{1|2}}$ is *highly skewed* unless our sample sizes are extremely large. So instead, we obtain a confidence interval for $\log(\frac{\pi_{1|1}}{\pi_{1|2}})$ based on $\log \frac{p_{1|1}}{p_{1|2}}$.

To derive the confidence interval, we use the ***delta method***.

The delta method for a function of a random variable:

Let T_n be a statistic, depending on a sample of size n. For large samples, suppose T_n is approximately normally distributed with mean θ and variance σ^2/n. Then as $n \to \infty$

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$$

Using a *Taylor series expansion* of $g(T_n)$ around θ, we can write

$$g(T_n) = g(\theta) + (T_n - \theta)g'(\theta) + (T_n - \theta)^2 \frac{g''(\theta)}{2} + ...$$

Thus we can get

$$\sqrt{n}\,[g(T_n) - g(\theta)] = \sqrt{n}\,(T_n - \theta)g'(\theta) + ...$$

and this implies that

$$\sqrt{n}\,[g(T_n) - g(\theta)]$$

has the same limiting distribution as

$$\sqrt{n}\,(T_n - \theta)g'(\theta)$$

$(T_n - \theta)$ converges in probability to 0 as $n \to \infty$ so we can write

$$\sqrt{n}\,[g(T_n) - g(\theta)] \xrightarrow{d} N\left(0, \sigma^2[g'(\theta)]^2\right)$$

In our case *we want a confidence interval for* $log\left(\frac{\pi_{1|1}}{\pi_{1|2}}\right)$ *for large samples*.

We begin with the point estimator of $log\left(\frac{\pi_{1|1}}{\pi_{1|2}}\right)$ which is

$$\log \frac{p_{1|1}}{p_{1|2}} = \log p_{1|1} - \log p_{1|2}.$$

Recall

$$\sqrt{n_{1+}}\,(p_{1|1} - \pi_{1|1}) \xrightarrow{d} N(0, \pi_{1|1}(1 - \pi_{1|1}))$$

so

$$\sqrt{n_{1+}}\,[\log(p_{1|1}) - \log(\pi_{1|1})] \xrightarrow{d} N\left(0, \frac{(1 - \pi_{1|1})}{\pi_{1|1}}\right)$$

because $\left[\frac{\partial \log(\pi_{1|1})}{\partial \pi_{1|1}}\right]^2 = \left[\frac{1}{\pi_{1|1}}\right]^2$.

Similarly

$$\sqrt{n_{2+}}\,[\log(p_{1|2}) - \log(\pi_{1|2})] \xrightarrow{d} N(0, \pi_{1|2}(1-\pi_{1|2})/\pi_{1|2})$$

so

$$[\log(p_{1|1}) - \log(p_{1|2})] - [\log(\pi_{1|1}) - \log(\pi_{1|2})] \xrightarrow{d} N\left(0, \frac{(1-\pi_{1|1})/\pi_{1|1}}{n_{1+}} + \frac{(1-\pi_{1|2})/\pi_{1|2}}{n_{2+}}\right)$$

So we have a $100(1-\alpha)\%$ confidence interval for $log(\frac{\pi_{1|1}}{\pi_{1|2}})$ is given by

$$\log\frac{p_{1|1}}{p_{1|2}} \pm z_{a/2}\sqrt{\frac{(1-p_{1|1})/p_{1|1}}{n_{1+}} + \frac{(1-p_{1|2})/p_{1|2}}{n_{2+}}}$$

For our example, the 95%*C.I.* for $log(\frac{\pi_{1|1}}{\pi_{1|2}})$ is

$$\log\left(\frac{0.0171}{0.0094}\right) \pm z_{a/2}\sqrt{\frac{(1-0.0171)/0.0171}{11034} + \frac{(1-0.0094)/(0.0094)}{11037}}$$

i.e.

$$0.598 \pm 1.96(0.121) \text{ or } \quad (0.360, 0.836).$$

Now taking antilogs, a 95%*C.I.* for the relative risk $\frac{\pi_{1|1}}{\pi_{1|2}}$ in our example is $(1.43, 2.31)$. This means that we are 95% confident in stating that, after 5 years, the proportion of MI cases for physicians taking a placebo every second day is between 1.43 and 2.31 times the proportion of MI cases for physicians taking a single aspirin every second day. Again it appears that taking aspirin is associated with a lower proportion of MI cases.

Note: There are times when we might want to estimate the ratio of the "failure" probabilities $\frac{\pi_{2|1}}{\pi_{2|2}}$ rather than the ratio of "success" probabilities $\frac{\pi_{1|1}}{\pi_{1|2}}$

Odds Ratio

Another measure of association in contingency tables is the *odds ratio* θ

Consider again our physician example. Within row 1, the odds that the response is in column 1 instead of column 2 is

$$\Omega_1 = \frac{\pi_{1|1}}{\pi_{2|1}}$$

Similarly within row 2, the corresponding odds ratio is

$$\Omega_2 = \frac{\pi_{1|2}}{\pi_{2|2}}$$

$\Omega_i > 1$ corresponds to the sitution where response 1 is more likely than response 2 in row i.

Within-row conditional distributions are identical *iff* $\Omega_1 = \Omega_2$. (i.e., the variables are independent).

The ratio of the two odds $\Omega_1$and Ω_2 is called the **odds (or cross product) ratio**

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\frac{\pi_{1|1}}{\pi_{2|1}}}{\frac{\pi_{1|2}}{\pi_{2|2}}} = \frac{\pi_{1|1}\pi_{2|2}}{\pi_{2|1}\pi_{1|2}}$$

$\theta = 1$ tells us that the response is not affected by the group.

We estimate the *population odds ratio* θ by the *sample odds ratio* $\hat{\theta}$ where

$$\hat{\theta} = \frac{p_{1|1}p_{2|2}}{p_{2|1}p_{1|2}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}.$$

For our example, the sample odds ratio is

$$\begin{aligned}\hat{\theta} &= [(189)(10933)]/[(10845)(104)] \\ &= 1.83\end{aligned}$$

meaning the odds of MI are 83% higher for physicians in the placebo group than in the aspirin group.

A $100(1-\alpha)\%$ *C.I.* for the population odds ratio θ is based on the sample odds ratio $\widehat{\theta}$ but again, since the sampling distribution of $\widehat{\theta}$ is *highly skewed* except for extremely large sample sizes, we first obtain a confidence interval for $\log(\theta)$. We base it on $\log\widehat{\theta}$ and use the delta method again.

$$\begin{aligned}\log(\theta) &= \log[\frac{\pi_{1|1}\pi_{2|2}}{\pi_{2|1}\pi_{1|2}}] \\ &= \log[\frac{\pi_{1|1}}{\pi_{2|1}}]-\log[\frac{\pi_{1|2}}{\pi_{2|2}}] \\ &= \log[\frac{\pi_{1|1}}{1-\pi_{1|1}}]-\log[\frac{\pi_{1|2}}{1-\pi_{1|2}}]\end{aligned}$$

and

$$\begin{aligned}\log\left(\widehat{\theta}\right) &= \log[\frac{p_{1|1}p_{2|2}}{p_{2|1}p_{1|2}}] \\ &= \log[\frac{p_{1|1}}{p_{2|1}}]-\log[\frac{p_{1|2}}{p_{2|2}}] \\ &= \log[\frac{p_{1|1}}{1-p_{1|1}}]-\log[\frac{p_{1|2}}{1-p_{1|2}}]\end{aligned}$$

Now using the delta method we can write that

$$\sqrt{n_{1+}}\,[\log(\frac{p_{1|1}}{1-p_{1|1}}) - \log(\frac{\pi_{1|1}}{1-\pi_{1|1}})] \xrightarrow{d} N(0, \frac{1}{\pi_{1|1}} + \frac{1}{1-\pi_{1|1}})$$

and

$$\sqrt{n_{2+}}\,[\log(\frac{p_{1|2}}{1-p_{1|2}}) - \log(\frac{\pi_{1|2}}{1-\pi_{1|2}})] \xrightarrow{d} N(0, \frac{1}{\pi_{1|2}} + \frac{1}{1-\pi_{1|2}})$$

giving

$$[\log(\widehat{\theta}) - \log(\theta)] \xrightarrow{d} N\left(0, \frac{1}{n_{1+}\pi_{1|1}} + \frac{1}{n_{1+}(1-\pi_{1|1})} + \frac{1}{n_{2+}\pi_{1|2}} + \frac{1}{n_{2+}(1-\pi_{1|2})}\right).$$

Now the variance

$$\frac{1}{n_{1+}\pi_{1|1}} + \frac{1}{n_{1+}(1-\pi_{1|1})} + \frac{1}{n_{2+}\pi_{1|2}} + \frac{1}{n_{2+}(1-\pi_{1|2})}$$

is estimated by

$$\begin{aligned}&\frac{1}{n_{1+}p_{1|1}} + \frac{1}{n_{1+}(1-p_{1|1})} + \frac{1}{n_{2+}p_{1|2}} + \frac{1}{n_{2+}(1-p_{1|2})} \\ &= \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\end{aligned}$$

Thus a $100(1-\alpha)\%$ *C.I.*. for $\log(\theta)$ is given by

$$\log(\widehat{\theta}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

For our example, a 95% C.I. for $\log(\theta)$ is given by

$$\log(1.83) \pm 1.96\sqrt{\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933}}$$

i.e. $0.605 \pm 1.96(0.123)$ or $(0.365, 0.846)$

Now taking antilogs, a 95% C.I. for θ is $(1.44, 2.33)$.)

We interpret this as: we are 95% confident that, after 5 years, the odds of MI for physicians taking a placebo every second day is between 1.44 and 2.33 times the odds of MI for physicians taking aspirin.

Relationship between Odds Ratio and Relative Risk

Since

$$\begin{aligned}\text{Odds Ratio} &= \frac{\pi_{1|1}\pi_{2|2}}{\pi_{2|1}\pi_{1|2}} \\ &= \frac{\pi_{1|1}(1-\pi_{1|2})}{\pi_{1|2}(1-\pi_{1|1})}\end{aligned}$$

and

$$\text{Relative Risk}(RR) = \frac{\pi_{1|1}}{\pi_{1|2}}$$

,

we have

$$\text{Odds Ratio} = RR\left(\frac{1-\pi_{1|2}}{1-\pi_{1|1}}\right)$$

i.e. when the probabilities of "success" for both groups (i.e.$\pi_{1|1}$ and $\pi_{1|2}$) are close to zero. , the odds ratio and the relative risk are similar. (This happens for our physician example and, in general, for a *rare condition*.)

SAS program for the physician example.

If the data is internal to the program:

```
data aspirin;
input Group $ MI $ count;
cards;
Placebo Yes 189
Placebo No 10845
Aspirin Yes 104
Aspirin No 10933
;
proc freq order = data;
tables GROUP*MI
/ chisq expected cellchi2 nocol nopct measures;
weight count;
run;
```

If the data is external to the program:

```
data aspirin;
infile 'k:/STAT5602/aspirin.txt';
input Group $ MI $ count;
proc freq order = data;
   tables GROUP*MI / chisq expected cellchi2 nocol nopct measures;
   weight count;
run;
```