Spring 2014
Stat 8004: Statistical Methods II
Lecture 6

Reference: Lachin (2002), Chapter 2.

1 Introduction to Biomedical Research

The goal of biomedical research is to advance our understanding of pathobiology or pathophysiology of diseases in man and the potential mechanisms for their treatment.

Type of Research:

- Basic Science Research ("Bench Science"): the study of genetic, biochemical, physiologic, and biological processes, (*e.g.*, study of genetic defects, metabolic pathways, pharmacology).
- Clinical Research ("Patient-oriented Research"): the study of clinical features of a population (*e.g.*, epidemiology, clinical outcomes, and health services research).
- Translational Research: taking information from basic science study and "translates" it for use in patient care.

There are many types research designs include those that compare two different independent groups:

- cross sectional study
- prospective cohort study
- retrospective study

Each concerned with comparing "risk" (usually defined/measured in terms of probability) in two independent groups.

Here are some definition of "risk".

Prevalence: Probability in the population, or proportion in a sample, with event/characteristic of interest in a cross-section of the population at a particular point in time. For example, the prevalence of adult onset type II diabetes as of 1980 was estimated to be approximately 6.8% of the US population.

Incidence: Probability in the population, or proportion in a sample, that acquire the event/characteristic of interest during a specified time interval for a randomly selected individual who was initially event/characteristic free. For example, it is estimated that the incidence of a new diagnosis of diabetes among adults in the US population is 2.42 new cases per 1,000 in the population per year.

Rate: the probability of an event among a group of people at risk for the event over a specific interval. For example, 5-year reoccurrence rate of cancer.

Here are some important elements involved in the biomedical (or other application fields) related statistical research.

1. What is the primary clinical outcome for the study?
    - What is the event?
    - Is it well defined conceptually?
    - Is it defined in terms of other variables?
2. What is the probability distribution that is relevant to the primary outcome?
3. What is the population?
4. Are there subsets of the populations that are defined by variables in the study that are important in estimating risk?
5. Are there experimentally assigned variables (design factors) that need to be taken into account in estimating risk?

2 Contingency Table

2.1 Definition

What is *contingency table*?

A contingency table is a data display good for

- Computing and comparing proportions
- Thinking about independence and patterns of dependence

- Computing conditional probabilities and odds
- Sampling schemes and sampling distributions

Example: Smoking and Lung Cancer.

		Lung Cancer?		
		Yes	No	Total
Smoking?	Yes	360	120	480
	No	95	105	200
	Total	455	225	680

Table 1: Contingency table for smoking and lung cancer data

More generally, we view "lung cancer" as the response, and "smoking" as the exposure. We are interested in the scientific question: Will the exposure increase/decrease the incidence rate of the particular response?

2.2 Model and Analysis

We can use different models to analyze a contingency table. Which model to use depends on how we collect the data, as well as the assumptions.

	Response		
Exposure	+	−	
1	a	b	n_1
2	c	d	n_2
	m_1	m_2	N

2.2.1 Model 1: Multinomial

Suppose it is decided in advance that we go out and sample 680 people at random. After the data are collected, we count the number of people falling in each category.

$$(a, b, c, d) \sim \text{Multinomial}(680; \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}),$$
$$\mathbb{P}(a, b, c, d) = \binom{N}{a, b, c, d} \pi_{11}^a \pi_{12}^b \pi_{21}^c \pi_{22}^d,$$

where $N = 680$.

2.2.2 Model 2: Poisson

Suppose we sample people for a fixed time period rather than to a fixed sample size.

$$\begin{aligned}\mathbb{P}(a,b,c,d) &= \mathbb{P}(a)\,\mathbb{P}(b)\,\mathbb{P}(c)\,\mathbb{P}(d)\\ &= \left(\frac{\lambda_{11}^a}{a!}e^{-\lambda_{11}}\right)\left(\frac{\lambda_{12}^b}{b!}e^{-\lambda_{12}}\right)\left(\frac{\lambda_{21}^c}{c!}e^{-\lambda_{21}}\right)\left(\frac{\lambda_{22}^d}{d!}e^{-\lambda_{22}}\right)\end{aligned}$$

2.2.3 Model 3: Product Binomial

Prospective Cohort Study. Suppose we randomly select 480 smokers and 200 non-smokers, and then see if they have lung cancer.

$$\begin{aligned}\mathbb{P}(a,b,c,d) &= \mathbb{P}(a,c \mid n_1,n_2)\\ &= \binom{n_1}{a}\pi_1^a(1-\pi_1)^{n_1-a}\binom{n_2}{c}\pi_2^c(1-\pi_2)^{n_2-c} \qquad (1)\end{aligned}$$

Retrospective/Case-Control Study. Suppose we randomly select 455 cancer patients and 225 non-cancer patients, and then check whether they are smokers.

$$\begin{aligned}\mathbb{P}(a,b,c,d) &= \mathbb{P}(a,b \mid m_1,m_2)\\ &= \binom{m_1}{a}\pi_1^a(1-\pi_1)^{m_1-a}\binom{m_2}{b}\pi_2^b(1-\pi_2)^{m_2-b}\end{aligned}$$

2.2.4 Model 4: Conditional Hypergeometric

Consider the *Prospective Design*. In Model (1), n_1 and n_2 are fixed. Under certain situations, we would also like to fix m_1. For example, we decided to terminate the study as long as we get m_1 cases. Then how to build up the model?

It is easy to see that $b = m_1 - a$. Plug it into (1), we will get

$$\mathbb{P}(a,m_1 \mid n_1,n_2,\pi_1,\pi_2) = \binom{n_1}{a}\binom{n_2}{m_1-a}\varphi^{x_1}(1-\pi_1)^{n_1}\pi_2^{m_1}(1-\pi_2)^{n_2-m_1}, \qquad (2)$$

where $\varphi = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ is the odds ratio. In (2), m_1 is still a random variable. To fix it, we can calculate the conditional probability $\mathbb{P}(a \mid m_1,n_1,n_2,\pi_1,\pi_2)$. Note that

$$\begin{aligned}\mathbb{P}(a \mid m_1,n_1,n_2,\pi_1,\pi_2) &= \frac{\mathbb{P}(a,m_1 \mid n_1,n_2,\pi_1,\pi_2)}{\mathbb{P}(m_1 \mid n_1,n_2,\pi_1,\pi_2)}\\ &= \frac{\mathbb{P}(a,m_1 \mid n_1,n_2,\pi_1,\pi_2)}{\sum_{i=a_l}^{a_u}\mathbb{P}(i,m_1 \mid n_1,n_2,\pi_1,\pi_2)},\end{aligned}$$

where $a_l = \max(0, m_1 - m_2)$ and $a_u = \min(m_1, n_1)$. It leads to the conditional hypergeometric model:

$$\mathbb{P}\left(a \mid m_1, n_1, n_2, \pi_1, \pi_2\right) = \mathbb{P}\left(a \mid m_1, n_1, n_2, \varphi\right) = \frac{\binom{n_1}{a}\binom{N-n_1}{m_1-a}\varphi^a}{\sum_{i=a_l}^{a_u}\binom{n_1}{i}\binom{N-n_1}{m_1-i}\varphi^i}. \tag{3}$$

This probability model only has one parameter, the odds ratio φ.

3 Measure of Risks

Type(θ)	Expression	Null Value
Risk difference (RD)	$\pi_1 - \pi_2$	0
Relative risk (RR)	π_1/π_2	1
$\log(RR)$	$\log(\pi_1) - \log(\pi_2)$	0
Odds ratio (OR)	$\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$	1
$\log(OR)$	$\log(\pi_1/(1-\pi_1)) - \log(\pi_2/(1-\pi_2))$	0

Table 2: Definition of Risks

Example: Smoking and Lung Cancer.

		Lung Cancer?		
		Yes	No	Total
Smoking?	Yes	$a = 360$	$b = 120$	$n_1 = 480$
	No	$c = 95$	$d = 105$	$n_2 = 200$
	Total	$m_1 = 455$	$m_2 = 225$	$N = 680$

Prospective Study. Group 1: smokers; Group 2: non-smokers. n_1 and n_2 are fixed numbers. Let

$$\pi_1 = \mathbb{P}\left(\text{Having lung cancer} \mid \text{Smoking}\right)$$
$$\pi_2 = \mathbb{P}\left(\text{Having lung cancer} \mid \text{Non-smoking}\right)$$

By comparing π_1 and π_2, we can know whether the lung cancer rate is the same between smokers and non-smokers.

How to estimate π_1 and π_2? Recall that for prospective study, n_1 and n_2 are fixed before the data are collected, and both smokers and non-smokers are randomized from the population of interest. Thus,

$$\hat{\pi}_1 = a/n_1 = 0.75; \quad \hat{\pi}_2 = c/n_2 = 0.475.$$

Based on Table 2, here are the estimators.

$$\widehat{RD} = \hat{\pi}_1 - \hat{\pi}_2 = 0.235.$$
$$\widehat{RR} = \hat{\pi}_1/\hat{\pi}_2 = 1.578 \Rightarrow \log(\widehat{RR}) = 0.457.$$
$$\widehat{OR} = \frac{\hat{\pi}_1(1-\hat{\pi}_2)}{\hat{\pi}_2(1-\hat{\pi}_1)} = 3.315 \Rightarrow \log(\widehat{OR}) = 1.199.$$

Retrospective Study. Group 1: Lung cancer subjects; Group 2: Lung cancer free subjects. m_1 and m_2 are fixed numbers. Let

$$\pi_1 = \mathbb{P}\left(\text{Smoking} \mid \text{Having lung cancer}\right)$$
$$\pi_2 = \mathbb{P}\left(\text{Smoking} \mid \text{Lung cancer free}\right)$$

By comparing π_1 and π_2, we can know whether the smoking rate is the same between lung cancer patients and non-lung cancer patients. Scientifically speaking, the problem is not as interesting as the previous one. Anyway, we can still estimate the corresponding risks.

$$\hat{\pi}_1 = a/m_1 \approx 0.791; \quad \hat{\pi}_2 = b/m_2 \approx 0.533.$$

Therefore,

$$\widehat{RD} = \hat{\pi}_1 - \hat{\pi}_2 = 0.258.$$
$$\widehat{RR} = \hat{\pi}_1/\hat{\pi}_2 = 1.483 \Rightarrow \log(\widehat{RR}) = 0.394.$$
$$\widehat{OR} = \frac{\hat{\pi}_1(1-\hat{\pi}_2)}{\hat{\pi}_2(1-\hat{\pi}_1)} = 3.315 \Rightarrow \log(\widehat{OR}) = 1.199.$$

It is easy to see that, due to different definitions of π_1 and π_2, the estimators of RD or RR are different for prospective and retrospective studies. However, OR is the same. It indicates that even if we are interested in the scientifically interesting question (as the one in the prospective study), we can still use retrospective design.

4 Inference

4.1 Large Sample Inference Under Product Binomial Model

4.1.1 Risk Difference

Estimate of $RD = \pi_1 - \pi_2$ is $\widehat{RD} = \hat{\pi}_1 - \hat{\pi}_2$.

$$\widehat{RD} \overset{d.}{\sim} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

We would like to test whether $\mathrm{H}_0: \ \pi_1 = \pi_2$. Equivalently, we test $\mathrm{H}_0: \ RD = 0$.

Under $\mathrm{H}_0: \ RD = 0$,

$$\widehat{RD} \overset{d.}{\sim} N\left(0, \hat{\sigma}_1^2\right),$$

where $\hat{\sigma}_1^2 = \hat{\pi}(1-\hat{\pi})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]$, where $\hat{\pi}$ is the proportion of response of interest in all the samples.

To control type I error at α, we reject H_0 iff $|\widehat{RD}| > z_{1-\alpha/2}\hat{\sigma}_1$.

Large sample $100(1-\alpha)\%$ CI for RD:

$$(\hat{\theta}_l, \hat{\theta}_u) = \hat{\theta} \pm z_{1-\alpha/2}\hat{\sigma}_1,$$

where $\hat{\theta} = \hat{\pi}_1 - \hat{\pi}_2$ and $\hat{\sigma}_2^2 = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}$.

Note that the estimators of the variance are different for hypothesis testing and confidence interval construction.

Example: Smoking and Lung Cancer.

Prospective study. RD is the risk difference of lung cancer between smoking and non-smoking patients.

Under $\mathrm{H}_0: \ RD = 0$, $\hat{\pi} = m_1/N = 0.67$.

$$\widehat{\mathrm{Var}}(\widehat{RD}) = \hat{\sigma}_1^2 = \hat{\pi}(1-\hat{\pi})\left[\frac{1}{n_1} + \frac{1}{n_2}\right] = 0.00157.$$

Hypothesis testing: $\widehat{RD}/\hat{\sigma}_1 = 6.944 > z_{0.975}$. Therefore, we reject H_0.

Under $\mathrm{H}_1: \ RD \neq 0$, $\hat{\pi}_1 = 0.75$, $\hat{\pi}_2 = 0.475$.

$$\widehat{\mathrm{Var}}(\widehat{RD}) = \hat{\sigma}_2^2 = \frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2} = 0.00164.$$

95% Confidence interval:

$$\hat{\theta}_l = \widehat{RD} - z_{0.975} \cdot \hat{\sigma}_2 = 0.20.$$
$$\hat{\theta}_u = \widehat{RD} + z_{0.975} \cdot \hat{\sigma}_2 = 0.35.$$

4.1.2 Relative Risk

Large sample distributional forms for $\widehat{RR}$ and $\widehat{OR}$ are better approximated using log transformation. For each group, Delta method implies

$$\begin{aligned}\text{Var}(\log(\hat{\pi}_i)) &\approx \left(\frac{\partial \log(x)}{\partial x}\right)^2\Bigg|_{x=\mathbb{E}(\hat{\pi}_i)} \text{Var}(\hat{\pi}_i) \\ &= \left(\frac{1}{\pi_i}\right)^2 \frac{\pi_i(1-\pi_i)}{n_i} \\ &= \frac{1-\pi_i}{n_i\pi_i}\end{aligned}$$

Besides, $\mathbb{E}(\log(\hat{\pi}_i)) \approx \log(\pi_i)$.

From Slutsky's Theorem,

$$\log(\hat{\pi}_i) \overset{d.}{\sim} N\left(\log(\pi_i), \frac{1-\pi_i}{\pi_i n_i}\right).$$

For prospective design, $\hat{\pi}_1$ and $\hat{\pi}_2$ are independent.

$$\begin{aligned}\mathbb{E}(\log(\widehat{RR})) &= \mathbb{E}(\log(\hat{\pi}_1) - \log(\hat{\pi}_2)) \approx \log(\pi_1) - \log(\pi_2) \\ \text{Var}(\log(\widehat{RR})) &= \text{Var}(\log(\hat{\pi}_1)) + \text{Var}(\log(\hat{\pi}_2)) = \frac{1-\pi_1}{\pi_1 n_1} + \frac{1-\pi_2}{\pi_2 n_2}\end{aligned}$$

For relative risk, $\text{H}_0: \pi_1 = \pi_2$ is equivalent to $\text{H}_0: RR = 1$.

Under H_0, $\pi_1 = \pi_2 = \pi$. Similar as before, we can estimate it with $\hat{\pi} = m_1/n$. Then

$$\hat{\sigma}_3^2 = \widehat{\text{Var}}(\log(\widehat{RR}))_{\text{H}_0} = \frac{n(1-\hat{\pi})}{n_1 n_2 \hat{\pi}}.$$

We reject H_0 if $\log(\widehat{RR})/\hat{\sigma}_3 > z_{1-\alpha/2}$.

Under H_1, we derive the large sample $(1-\alpha)$ CI for $\log(RR)$:

$$(\hat{\theta}_l, \hat{\theta}_u) = \log(\widehat{RR}) \pm z_{1-\alpha/2}\hat{\sigma}_4,$$

where $\hat{\theta} = \log(\hat{\pi}_1) - \log(\hat{\pi}_2)$, and

$$\hat{\sigma}_4^2 = \frac{1-\hat{\pi}_1}{m_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{m_2\hat{\pi}_2}.$$

Asymmetric confidence interval for RR are: $(\widehat{RR}_l, \widehat{RR}_u) = (\exp(\hat{\theta}_l), \exp(\hat{\theta}_u))$.

Example: Smoking and Lung Cancer

Prospective study. RR is the relative risk of lung cancer between the smoking and non-smoking group.

Under H_0, $\hat{\pi} = 0.67$, $\hat{\sigma}_3 = 0.059$. $\log(\widehat{RR})/\hat{\sigma}_3 = 7.71 > z_{0.975}$. Therefore, we reject H_0.

We first get 95% CI for $\log(RR)$, which is $(0.30, 0.61)$, and therefore 95% CI for RR is $(1.35, 1.84)$.

4.2 Odds Ratio

By similar procedures above, we can have

$$\begin{aligned}
\hat{\varphi} = \widehat{OR} &= \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{ad}{bc} \\
\log(\hat{\varphi}) = \log(\widehat{OR}) &= \log(a) + \log(d) - \log(b) - \log(c) \\
\mathrm{Var}(\log(\widehat{OR})) &= \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)}
\end{aligned}$$

To test $\mathrm{H}_0: \ \pi_1 = \pi_2$ is equivalent to test $\mathrm{H}_0: \ OR = 1$.

Under H_0, $\hat{\pi}_1 = \hat{\pi}_2 = \hat{\pi} = m_1/n$. Then

$$\hat{\sigma}_5^2 = \widehat{\mathrm{Var}}(\log(\widehat{OR}))_{\mathrm{H}_0} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{1}{\hat{\pi}(1-\hat{\pi})}.$$

We reject H_0 if $\log(\widehat{OR})/\hat{\sigma}_5 > z_{1-\alpha/2}$.

Under H_1,

$$\hat{\sigma}_6^2 = \widehat{\mathrm{Var}}(\log(\widehat{OR})) = 1/(n_1\hat{\pi}_1(1-\hat{\pi}_1)) + 1/(n_2\hat{\pi}_2(1-\hat{\pi}_2)) = 1/a + 1/b + 1/c + 1/d.$$

Large sample $(1-\alpha)$ CI for $\log(OR)$ is

$$(\hat{\theta}_l, \hat{\theta}_u) = \log(\widehat{OR}) \pm z_{1-\alpha/2}\hat{\sigma}_6,$$

Asymmetric confidence interval for RR are: $(\widehat{OR}_l, \widehat{OR}_u) = (\exp(\hat{\theta}_l), \exp(\hat{\theta}_u))$.

Example: Smoking and Lung Cancer.

To test H_0, $\hat{\sigma}_5 = 0.177$. $\log(\widehat{OR})/\hat{\sigma}_5 = 6.79 > z_{0.975}$. Thus, we reject H_0.

95% CI for $\log(OR)$ is $(0.85, 1.54)$, for OR is $(2.35, 4.69)$.

4.3 Exact Inference Under Conditional Hypergeometric Model

To get the exact inference, we can based on the conditional hypergeometric model. It can be shown that (3), the larger φ, the most likely that a will be large. In other words, for large φ, the extreme situations are those with small a, and for small φ, the extreme situations are those with large a.

We first derive the confidence interval. $100(1-\alpha)\%$ confidence limits $(\hat{\varphi}_l, \hat{\varphi}_u)$:

- $\hat{\varphi}_l$ is the largest number that satisfies

$$\frac{\alpha}{2} \geq \sum_{x=a}^{\min(m_1,n_1)} \mathbb{P}\left(x \mid m_1, n_1, n_2, \hat{\varphi}_l\right). \tag{4}$$

- $\hat{\varphi}_u$ is the smallest number that satisfies

$$\frac{\alpha}{2} \geq \sum_{x=\max(0,m_1-n_2)}^{a} \mathbb{P}\left(x \mid m_1, n_1, n_2, \hat{\varphi}_u\right). \tag{5}$$

Some special situation

- If $a = \max(0, m_1 - n_2)$, (4) has no solution $\Rightarrow$ lower limit $= 0$.
- If $a = \min(m_1, n_1)$, (5) has no solution $\Rightarrow$ upper limit $= \infty$.

Now we discuss how to test $\mathrm{H}_0 : \ OR = \varphi = 1$ based on the conditional hypergeometric distribution.

Let $\varphi_0 = 1$. Under H_0, the probability distribution of a is

$$\mathbb{P}\left(a \mid m_1, n_1, n_2, \varphi_0\right) = \frac{\binom{n_1}{a}\binom{N-n_1}{m_1-a}}{\sum_{i=a_l}^{a_u}\binom{n_1}{i}\binom{N-n_1}{m_1-i}}.$$

It can be shown that the denominator equals $\binom{N}{m_1}$ so that the distribution reduces to

$$\mathbb{P}\left(a \mid m_1, n_1, n_2, \varphi_0\right) = \frac{\binom{n_1}{a}\binom{n_2}{m_1-a}}{\binom{N}{m_1}} = \frac{\binom{m_1}{a}\binom{m_2}{n_1-a}}{\binom{N}{n_1}} = \frac{n_1!n_2!m_1!m_2!}{n!a!b!c!d!}.$$

To test H_0 *vs.* $\mathrm{H}_1 : \pi_1 < \pi_2$, the exact one-sided left tailed p-value is

$$p_l = \sum_{x=a_l}^{a} \mathbb{P}\left(x \mid n_1, n_2, m_1, \phi_0\right).$$

To test H_0 *vs.* $\mathrm{H}_1 : \pi_1 > \pi_2$, the exact one-sided right tailed p-value is

$$p_u = \sum_{x=a}^{a_u} \mathbb{P}\left(x \mid n_1, n_2, m_1, \phi_0\right).$$

For two sided tests ($\mathrm{H}_1 : \pi_1 \neq \pi_2$), the exact p-value is $p = p_l + p_u$.

The test is called *Fisher-Irwin Exact Test.*

4.4 Other Large Sample Hypothesis Testing Methods

4.4.1 Pearson Contingency Chi-Square Test

Pearson Chi-square test is one of the most commonly used test for $R \times C$ contingency table.

$$X_P^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}},$$

where O_{ij} is the observed frequency in the ith row and jth column and $\hat{E}_{ij}$ is the estimated expected frequency under the null hypothesis.

We would like to know whether the exposure is associated with the outcome. How to set up the hypothesis?

$$\mathrm{H}_0 : \ \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}.$$

Under the null,

$$E_{ij} = \mathbb{E}\left(O_{ij}\right) = n\pi_{ij} = n\pi_{i\cdot}\pi_{\cdot j}.$$

Substituting the sample estimators: $\hat{\pi}_{i\cdot} = n_{i\cdot}/n$ and $\hat{\pi}_{\cdot j} = n_{\cdot j}/n$ yields the estimated expected frequencies:

$$\hat{E}_{ij} = n_{i\cdot}n_{\cdot j}/n.$$

Under the null, asymptotically $X_P^2 \xrightarrow{d.} \chi^2_{(R-1)(C-1)}$.

For a 2×2 table, $|O_{ij} - \hat{E}_{ij}|$ is constant for all celles of the table. Thus,

$$X_P^2 = \frac{(ad - bc)^2 N}{n_1 n_2 m_1 m_2},$$

which is asymptotically distributed as $\chi^2(1)$ under H_0.

Example: Smoking and Lung Cancer. $X_P^2 = 46.99$, $df = 1$, the $p - value = 7.14E - 12$. We reject H_0 under the level 0.05.

4.5 Conditional Mantel-Haenszel Test

Consider the Hypergeometric distribution. We discussed the exact test just now. We can also build up large sample test under this model.

Under $\mathrm{H}_0: \ \varphi = OR = 1$,

$$\mathbb{E}(a) = \frac{n_1 m_1}{n}$$
$$\widehat{\mathrm{Var}}(a) = \frac{n_1 n_2 m_1 m_2}{n^2(n-1)}$$

Conditional Mantel-Haeszel Test:

$$X_c^2 = \frac{(a - \mathbb{E}(a))^2}{\widehat{\mathrm{Var}}(a)},$$

and the corresponding z-statistics is

$$Z_c = \frac{a - \mathbb{E}(a)}{\sqrt{\widehat{\mathrm{Var}}(a)}}.$$

Under H_0, $X_c^2 \xrightarrow{d.} \chi^2(1)$ and $Z_c \xrightarrow{d.} N(0,1)$.

Example: Smoking and Lung Cancer. $Z_c = 6.93 > z_{0.975}$, and therefore, we reject H_0.

4.6 Cochran's Test

Consider the Product Binomial Model. Here is another option to test $\mathrm{H}_0: \ \pi_1 = \pi_2 = \pi$. Instead of testing on the risks, we develop tests based on $[a - \mathbb{E}(a)]$.

Under the Product Binomial Distribution,

$$\mathbb{E}(a) = n_1 \pi_1$$
$$\mathrm{Var}(a) = n_1 \pi_1 (1 - \pi_1).$$

They can be reasonably estimated as

$$\widehat{\mathbb{E}}(a) = n_1 \hat{\pi}_1 = n_1 m_1 / n$$
$$\widehat{\mathrm{Var}}(a) = n_1 m_1 m_2 / n^2$$

Likewise,

$$\mathrm{Var}(c) = n_1 \pi_2 (1 - \pi_2).$$
$$\widehat{\mathrm{Var}}(c) = n_2 m_1 m_2 / n^2$$

Based on these results, under $\mathrm{H}_0: \ \pi = \pi_1 = \pi_2$.

$$V_u = \mathrm{Var}[a - \hat{E}(a)] = \frac{n_2^2 \mathrm{Var}(a) + n_1^2 \mathrm{Var}(c)}{n^2} = \frac{n_1 n_2 \pi (1 - \pi)}{n},$$

which can be estimated as

$$\hat{V}_u = \widehat{\text{Var}}[a - \widehat{\mathbb{E}}(a)] = \frac{n_1 n_2 \hat{\pi}(1-\hat{\pi})}{n} = \frac{n_1 n_2 m_1 m_2}{n^3}.$$

The Cochran's Chi-square test statistic for 2×2 table is

$$X_u^2 = \frac{[a - \widehat{\mathbb{E}}(a)]^2}{\hat{V}_u},$$

which can be shown to be equal to X_P^2. The corresponding z-statistic is

$$Z_u = \frac{a - \widehat{\mathbb{E}}(a)}{\sqrt{\hat{V}_u}}.$$

Under H_0, asymptotically X_u^2 follows $\chi^2(1)$ and Z_u follows $N(0,1)$.

4.7 Likelihood Ratio Test

Like the Pearson test, this test tests the independence of the row and the column factors. It can be adopted to $R \times C$ contingency table.

$$G^2 = 2 \sum_{i=1}^{R} \sum_{j=1}^{C} O_{ij} \log \left(\frac{O_{ij}}{\hat{E}_{ij}} \right),$$

where O_{ij} and $\hat{E}_{ij}$ are observed and estimated expected frequencies defined above. Asymptotically, under $\text{H}_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$, $G^2 \sim \chi^2((R-1)(C-1))$.

Example: Smoking and Lung Cancer. $G^2 = 46.73 > z_{\chi^2(1),0.95}$. And therefore, we reject H_0.