Contiingency Tables

2×2

Demetris Athienitis

Notation

Commonly the overall sample size, n, is fixed by design and sometimes the row totals are fixed by design.

- Joint probabilities are no longer useful
- Can use the binomial distribution within each row

Notation $\pi_1 = \pi_{1|1} = P(Y = 1|X = 1)$ and $\pi_2 = \pi_{1|2} = P(Y = 1|X = 2)$.

		Y: 1	Y: 2
X	1	π_1	$1 - \pi_1$
	2	π_2	$1 - \pi_2$

Notation

Commonly the overall sample size, n, is fixed by design and sometimes the row totals are fixed by design.

- Joint probabilities are no longer useful
- Can use the binomial distribution within each row

Notation $\pi_1 = \pi_{1|1} = P(Y = 1|X = 1)$ and $\pi_2 = \pi_{1|2} = P(Y = 1|X = 2)$.

		Y = 1	Y = 2
X	1	π_1	$1 - \pi_1$
	2	π_2	$1 - \pi_2$

Assuming the two levels of X are independent, we use the same formula from your introductory statistics class to create the $100(1-\alpha)\%$ CI on $\pi_1 - \pi_2$

$$p_1 - p_2 \mp z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Difference of Proportions

Assuming the two levels of X are independent, we use the same formula from your introductory statistics class to create the $100(1-\alpha)\%$ CI on $\pi_1 - \pi_2$

$$p_1 - p_2 \mp z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Example (Physicians' Health Study ctd)

Look at the probability of heart attack given the treatment group.

Group	MI	
	Yes	No
Placebo	189	10845
Aspirin	104	10933

$\rightarrow$

Group	MI		
	Yes	No	Total
Placebo	0.017	0.983	1
Aspirin	0.009	0.991	1

A 95% CI for $\pi_1 - \pi_2$

$$0.017 - 0.009 \mp 1.96\sqrt{\frac{0.017(0.983)}{11034} + \frac{0.009(0.991)}{11037}} \rightarrow (0.005, 0.011)$$

Those on placebo have a higher chance of having an MI by at least 0.005 and at most 0.011 (with the point eastimate of 0.008).

Section 2

Relative Risk

Definition (Relative Risk)

Relative Risk (R.R.) is defined as

$$R.R. = \frac{\pi_1}{\pi_2}$$

Example

From the MI example, R.R.=1.82. Hence, the sample proportion of heart attacks was 82% higher for placebo group, which better portrays the difference (compared to difference of proportions).

R.R. CI

Using the *Delta Method*, a $100(1-\alpha)\%$ CI on $\log(\pi_1/\pi_2)$ is

$$\log\left(\frac{p_1}{p_2}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1-p_1}{(n_{1+})p_1} + \frac{1-p_2}{(n_{2+})p_2}} \rightarrow (L, U)$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Consequently, $100(1-\alpha)$ CI on π_1/π_2 is (e^L, e^U). If 1 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Example

From the MI example, a 95% CI for $\log(\pi_1/\pi_2)$ ends up being $(0.3571, 0.8406)$ and hence for π_1/π_2

$$(e^{0.3571}, e^{0.8406}) \rightarrow (1.43, 2.31)$$

R.R. CI

Using the *Delta Method*, a $100(1-\alpha)\%$ CI on $\log(\pi_1/\pi_2)$ is

$$\log\left(\frac{p_1}{p_2}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1-p_1}{(n_{1+})p_1} + \frac{1-p_2}{(n_{2+})p_2}} \to (L, U)$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Consequently, $100(1-\alpha)$ CI on π_1/π_2 is (e^L, e^U). If 1 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Example

From the MI example, a 95% CI for $\log(\pi_1/\pi_2)$ ends up being $(0.3571, 0.8406)$ and hence for π_1/π_2

$$(e^{0.3571}, e^{0.8406}) \to (1.43, 2.31)$$

R.R. CI

Using the *Delta Method*, a $100(1-\alpha)\%$ CI on $\log(\pi_1/\pi_2)$ is

$$\log\left(\frac{p_1}{p_2}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1-p_1}{(n_{1+})p_1} + \frac{1-p_2}{(n_{2+})p_2}} \rightarrow (L, U)$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Consequently, $100(1-\alpha)$ CI on π_1/π_2 is (e^L, e^U). If 1 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Example

From the MI example, a 95% CI for $\log(\pi_1/\pi_2)$ ends up being $(0.3571, 0.8406)$ and hence for π_1/π_2

$$\left(e^{0.3571}, e^{0.8406}\right) \rightarrow (1.43, 2.31)$$

Odds Ratio

Redefine $Y = 1$ as a success and $Y = 2$ as a failure, the odds of success are

$$\text{odds(S)} = \begin{cases} \frac{\pi_1}{1-\pi_1} & X = 1 \\ \frac{\pi_2}{1-\pi_2} & X = 2 \end{cases}$$

Definition (Odds Ratio)

The Odds Ratio (O.R.) is the ratio of the odds of $Y = 1|X = 1$ to that of $Y = 1|X = 2$.

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Odds Ratio

Redefine $Y = 1$ as a success and $Y = 2$ as a failure, the odds of success are

$$\text{odds(S)} = \begin{cases} \frac{\pi_1}{1-\pi_1} & X = 1 \\ \frac{\pi_2}{1-\pi_2} & X = 2 \end{cases}$$

Definition (Odds Ratio)

The Odds Ratio (O.R.) is the ratio of the odds of $Y = 1|X = 1$ to that of $Y = 1|X = 2$.

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

Example

Group	MI	
	Yes	No
Placebo	189	10845
Aspirin	104	10933

$\rightarrow$

Group	MI		
	Yes	No	Total
Placebo	0.017	0.983	1
Aspirin	0.009	0.991	1

$$\hat{\theta} = \frac{0.0171/0.9829}{0.0094/0.9906} = \frac{189 \times 10933}{104 \times 10845} = 1.83$$

The estimated odds of heart attack in placebo group are 1.83 times the odds of heart attack in the aspirin group.

O.R. CI

Using the Delta Method, the $100(1-\alpha)\%$ C.I. on $\log(\theta)$ is

$$\log\left(\hat{\theta}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} \rightarrow (L, U)$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Consequently, $100(1-\alpha)\%$ CI on θ is (e^L, e^U). If 1 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Example

From the MI example, a 95% CI for $\log(\theta)$

$$\log(1.83) \mp 1.96\sqrt{1/189 + 1/10845 + 1/104 + 1/10933} \rightarrow (0.365, 0.846)$$

and hence for θ, (1.44,2.33).

O.R. CI

Using the Delta Method, the $100(1-\alpha)\%$ C.I. on $\log(\theta)$ is

$$\log\left(\hat{\theta}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} \rightarrow (L, U)$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Consequently, $100(1-\alpha)\%$ CI on θ is (e^L, e^U). If 1 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Example

From the MI example, a 95% CI for $\log(\theta)$

$$\log(1.83) \mp 1.96\sqrt{1/189 + 1/10845 + 1/104 + 1/10933} \rightarrow (0.365, 0.846)$$

and hence for θ, (1.44,2.33).

O.R. CI

Using the Delta Method, the $100(1-\alpha)\%$ C.I. on $\log(\theta)$ is

$$\log\left(\hat{\theta}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}+\frac{1}{n_{22}}} \rightarrow (L, U)$$

If 0 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Consequently, $100(1-\alpha)\%$ CI on θ is (e^L, e^U). If 1 is in the CI that would imply $\pi_1 = \pi_2$, i.e. independence.

Example

From the MI example, a 95% CI for $\log(\theta)$

$$\log(1.83) \mp 1.96\sqrt{1/189+1/10845+1/104+1/10933} \rightarrow (0.365, 0.846)$$

and hence for θ, (1.44,2.33).

O.R. Properties

- If $1 < \theta < \infty$, the odds of success are *higher* in row 1 than in row 2
- If $0 < \theta < 1$, a success is *less* likely in row 1 than in row 2
- $\theta = 1 \Leftrightarrow \log(\theta) = 0$. This also implies $\pi_1 = \pi_2$, hence independence
- If rows are interchanged (or columns, but not both), $\theta \to 1/\theta$
- O.R. is valid for retrospective studies while R.R. and differencing are not. In retrospective studies, sampling is done within levels of Y, not to Y, and we cannot estimate $P(Y|X)$. We can estimate $P(X|Y)$ and hence θ, as θ treats rows and columns symmetrically

$$\begin{aligned}\theta &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)}\\ &= \cdots\\ &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}\end{aligned}$$

O.R. Properties

- If $1 < \theta < \infty$, the odds of success are *higher* in row 1 than in row 2
- If $0 < \theta < 1$, a success is *less* likely in row 1 than in row 2
- $\theta = 1 \Leftrightarrow \log(\theta) = 0$. This also implies $\pi_1 = \pi_2$, hence independence
- If rows are interchanged (or columns, but not both), $\theta \to 1/\theta$
- O.R. is valid for retrospective studies while R.R. and differencing are not. In retrospective studies, sampling is done within levels of Y, not to Y, and we cannot estimate $P(Y|X)$. We can estimate $P(X|Y)$ and hence θ, as θ treats rows and columns symmetrically

$$\begin{aligned}\theta &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)}\\ &= \cdots \\ &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}\end{aligned}$$

O.R. Properties

- If $1 < \theta < \infty$, the odds of success are *higher* in row 1 than in row 2
- If $0 < \theta < 1$, a success is *less* likely in row 1 than in row 2
- $\theta = 1 \Leftrightarrow \log(\theta) = 0$. This also implies $\pi_1 = \pi_2$, hence independence
- If rows are interchanged (or columns, but not both), $\theta \to 1/\theta$
- O.R. is valid for retrospective studies while R.R. and differencing are not. In retrospective studies, sampling is done within levels of Y, not to Y, and we cannot estimate $P(Y|X)$. We can estimate $P(X|Y)$ and hence θ, as θ treats rows and columns symmetrically

$$\begin{aligned}\theta &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)} \\ &= \cdots \\ &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}\end{aligned}$$

O.R. Properties

- If $1 < \theta < \infty$, the odds of success are *higher* in row 1 than in row 2
- If $0 < \theta < 1$, a success is *less* likely in row 1 than in row 2
- $\theta = 1 \Leftrightarrow \log(\theta) = 0$. This also implies $\pi_1 = \pi_2$, hence independence
- If rows are interchanged (or columns, but not both), $\theta \rightarrow 1/\theta$
- O.R. is valid for retrospective studies while R.R. and differencing are not. In retrospective studies, sampling is done within levels of Y, not to Y, and we cannot estimate $P(Y|X)$. We can estimate $P(X|Y)$ and hence θ, as θ treats rows and columns symmetrically

$$\begin{aligned}\theta &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)} \\ &= \cdots \\ &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}\end{aligned}$$

O.R. Properties

- If $1 < \theta < \infty$, the odds of success are *higher* in row 1 than in row 2
- If $0 < \theta < 1$, a success is *less* likely in row 1 than in row 2
- $\theta = 1 \Leftrightarrow \log(\theta) = 0$. This also implies $\pi_1 = \pi_2$, hence independence
- If rows are interchanged (or columns, but not both), $\theta \to 1/\theta$
- O.R. is valid for retrospective studies while R.R. and differencing are not. In retrospective studies, sampling is done within levels of Y, not to Y, and we cannot estimate $P(Y|X)$. We can estimate $P(X|Y)$ and hence θ, as θ treats rows and columns symmetrically

$$\begin{aligned}\theta &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)} \\ &= \cdots \\ &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}\end{aligned}$$

Example (Case-control study (Doll and Hill 1950))

- X = smoked at least 1 cigarette per day for at least 1 year
- $Y = 1$ for lung cancer, 0 otherwise

Smoked	Cancer	
	Yes	No
Yes	688	650
No	21	59
Total	709	709

Case-control study because they found 709 without lung cancer and then 709 with lung cancer; and *then* looked at whether they smoked or not.

$$\hat{\theta} = \frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{21 \times 650} = 2.97$$

Odds of lung cancer for smokers is estimated to be about 3 times the odds for non smokers.

Example (Case-control study (Doll and Hill 1950))

- X = smoked at least 1 cigarette per day for at least 1 year
- $Y = 1$ for lung cancer, 0 otherwise

Smoked	Cancer	
	Yes	No
Yes	688	650
No	21	59
Total	709	709

Case-control study because they found 709 without lung cancer and then 709 with lung cancer; and *then* looked at whether they smoked or not.

$$\hat{\theta} = \frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{21 \times 650} = 2.97$$

Odds of lung cancer for smokers is estimated to be about 3 times the odds for non smokers.

Example (Case-control study (Doll and Hill 1950))

- X = smoked at least 1 cigarette per day for at least 1 year
- $Y = 1$ for lung cancer, 0 otherwise

Smoked	Cancer	
	Yes	No
Yes	688	650
No	21	59
Total	709	709

Case-control study because they found 709 without lung cancer and then 709 with lung cancer; and *then* looked at whether they smoked or not.

$$\hat{\theta} = \frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{21 \times 650} = 2.97$$

Odds of lung cancer for smokers is estimated to be about 3 times the odds for non smokers.

Remark

- When any values $n_{ij} \approx 0$, it is best to use $\{n_{ij} + 0.5\}$
- When π_1 and π_2 are close to zero then O.R. $\approx$ R.R.

Remark

- When any values $n_{ij} \approx 0$, it is best to use $\{n_{ij} + 0.5\}$
- When π_1 and π_2 are close to zero then O.R. $\approx$ R.R.

We learned

3 ways of comparing whether

$$P(Y = 1|X = 1) = P(Y = 1|X = 2)$$

in a 2×2 table, using

- Difference of proportions
- Relative Risk
- Odds Ratio