Categorical Data Analysis
Lecture 3 & 4

Graduate School of Advanced Science and Engineering
Rei Monden

2024.12.12

Contents of this lecture

Today's contents

- Contingency tables and probabilities
- Comparing proportions in 2×2 contingency tables
  - Difference of proportions
  - Ratio of proportions (relative risk)
  - The odds ratio
- Chi-squared tests of independence

Contingency tables and probabilities

Contingency tables

Contingency tables cross-classify the counts of two or more categorical variables.

Here we limit to two categorical variables X (with r categories) and Y (with c categories).

	Y categories			
X categories	Y_1	Y_2	$\cdots$	Y_c
X_1	n_{11}	n_{12}	$\cdots$	n_{1c}
X_2	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

An $r \times c$ *two-way* contingency table has r *rows* and c *columns*, in a total of rc *cells*.

Contingency tables – three types of probabilities

There are three types of probabilities:

- Joint probability.
- Marginal probability.
- Conditional probability.

Let's quickly go through each type.

Always make sure you use the right one, depending on the intended purpose!

Contingency tables – Joint probability

	Y categories			
X categories	Y_1	Y_2	$\cdots$	Y_c
X_1	n_{11}	n_{12}	$\cdots$	n_{1c}
X_2	n_{21}	n_{22}	$\cdots$	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

Joint probability = cell probability:

$$\pi_{ij} = P(X = i, Y = j).$$

Estimated by

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n},$$

where $n = \sum n_{ij}$ is the sum of all counts (total sample size).

Contingency tables – Marginal probability

	Y categories				
X categories	Y_1	Y_2	$\cdots$	Y_c	
X_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
X_2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

Marginal probability = row or column probabilities:

$$\pi_{i+} = P(X = i) = \sum_j P(X = i, Y = j)$$

$$\pi_{+j} = P(Y = j) = \sum_i P(X = i, Y = j)$$

Estimated by

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n}, \quad \hat{\pi}_{+j} = \frac{n_{+j}}{n}.$$

Contingency tables – Conditional probability

	Y categories				
X categories	Y_1	Y_2	$\cdots$	Y_c	
X_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
X_2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_i	n_{i1}	n_{i2}	$\cdots$	n_{ic}	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}

Conditional probability for Y, given X = focus on row i:

$$P(Y = j|X = i) = \frac{\pi_{ij}}{\pi_{i+}}.$$

Estimated by

$$\frac{n_{ij}}{n_{i+}}.$$

Contingency tables – Example

Gender	Belief in Afterlife: Yes	Belief in Afterlife: No/Undecided	Total
Females	1230	357	1587
Males	859	413	1272
Total	2089	770	2859

- $\hat{\pi}_{12} = \frac{357}{2859} = .12$
- $\hat{\pi}_{+2} = \frac{770}{2859} = .27$
- $P(\text{Yes}|\text{Males}) = \frac{859}{1272} = .68.$

Contingency tables and independence

Categorical variables X and Y are said to be statistically independent when all joint probabilites equal the product of their marginal probabilities:

$$\underbrace{P(X=i, Y=j)}_{\pi_{ij}} = \underbrace{P(X=i)}_{\pi_{i+}} \underbrace{P(Y=j)}_{\pi_{+j}},$$

for all $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

Equivalently, X and Y are statistically equivalent iff

$$P(Y=j|X=i) = P(Y=j),$$

i.e., if all conditional probabilities equal the corresponding marginal probabilities.

Contingency tables and independence

Gender	Belief in Afterlife: Yes	Belief in Afterlife: No/Undecided	Total
Females	1230	357	1587
Males	859	413	1272
Total	2089	770	2859

Since

$$P(\text{Females}, \text{Yes}) = \frac{1230}{2859} = .430$$

and

$$P(\text{Females})P(\text{Yes}) = \frac{1587}{2859} \times \frac{2089}{2859} = .406$$

are different, *at least in the sample*, we conclude that 'Gender' and 'Belief in Afterlife' are statistically dependent.

Contingency tables and independence

Gender	Belief in Afterlife: Yes	Belief in Afterlife: No/Undecided	Total
Females	1230	357	1587
Males	859	413	1272
Total	2089	770	2859

Since

$$P(\text{Yes}|\text{Females}) = \frac{1230}{1587} = .775$$

and

$$P(\text{Yes}) = \frac{2089}{2859} = .731$$

are different, *at least in the sample*, we conclude that 'Gender' and 'Belief in Afterlife' are statistically dependent.

Comparing proportions in 2×2 contingency tables

Comparing proportions in 2×2 contingency tables

	Y categories	
X categories	Y_1	Y_2
X_1	n_{11}	n_{12}
X_2	n_{21}	n_{22}

- *Groups to be compared:*
  X_1 and X_2 (rows).
- *Dependent variable:*
  Binary variable Y (say, Y_1 = success; Y_2 = failure).
- *What we compare:*
  Proportions of success among the two groups:

$$\underbrace{\pi_1 = P(Y = 1|X = 1)}_{\text{Proportion of success in Group 1}} \qquad \text{versus} \qquad \underbrace{\pi_2 = P(Y = 1|X = 2)}_{\text{Proportion of success in Group 2}}$$

Comparing proportions in 2×2 contingency tables

There are three main strategies to compare the two proportions of success:

- Difference of proportions;
- Ratio of proportions (relative risk);
- The odds ratio.

Let's study one at a time.

Difference of proportions

Difference of proportions

The idea is to estimate

$$\pi_1 - \pi_2,$$

which compares the success probabilities for the two groups.

$(\pi_1 - \pi_2)$ is:

- Between -1 and $+1$;
- Exactly 0 when both proportions coincide (i.e., when X and Y are independent).

Difference of proportions

	Y categories		
X categories	Y_1	Y_2	Total
X_1	n_{11}	n_{12}	n_1
X_2	n_{21}	n_{22}	n_2
			n

$n_1 = n_{1+}$; $n_2 = n_{2+}$

The sample estimate of $(\pi_1 - \pi_2)$ is really simple:

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{n_{11}}{n_1} - \frac{n_{21}}{n_2}.$$

Difference of proportions

X categories	Y categories: Y_1	Y_2	Total
X_1	n_{11}	n_{12}	n_1
X_2	n_{21}	n_{22}	n_2
			n

$n_1 = n_{1+}$; $n_2 = n_{2+}$

The $100(1-\alpha)\%$ Wald confidence interval (CI) for $(\pi_1 - \pi_2)$ is:

$$\boxed{(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}(SE)},$$

with

$$SE = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

and $z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ quantile from $\mathcal{N}(0,1)$.

Difference of proportions

$$\boxed{(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}(SE)}$$

This CI is valid for large samples only.

For small samples, especially when π_1 and π_2 are close to 0 or 1, Wald's CI performs poorly.

In such cases, rely on:

- Score CI instead of Wald CI (we will compute it in R),

or

- Add 1 to each of the four cells before applying Wald's CI (Agresti-Caffo CI).

Difference of proportions - Example

Group	Myocardial infarction Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
			22071

95% Wald CI for $(\pi_1 - \pi_2)$:

- $\hat{\pi}_1 = \frac{189}{11034} = .0171$
- $\hat{\pi}_2 = \frac{104}{11037} = .0094$

$$\begin{aligned}
\text{95\% CI} &= (\hat{\pi}_1 - \hat{\pi}_2) \pm 1.96\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \\
&= (.0171 - .0094) \pm 1.96\sqrt{\frac{.0171(.9829)}{11034} + \frac{.0094(.9906)}{11037}} \\
&= (.0047, .0107).
\end{aligned}$$

Difference of proportions - Example

$$\boxed{95\% \text{ CI} = (.0047, .0107)}$$

This interval is entirely positive (i.e., it leaves 0 out).

Thus, there is evidence against $\mathcal{H}_0 : \pi_1 = \pi_2$.

Conclusion:

We decide to retain $\pi_1 > \pi_2$ and conclude that *evidence suggests* that taking aspirin diminishes the risk of heart attack.

Difference of proportions - Example

For these data, the sample size is large.

But, the group proportions are rather small (close to 0):

$$\hat{\pi}_1 = .0171 \qquad , \qquad \hat{\pi}_2 = .0094.$$

We could instead use the Agresti-Caffo CI:

$$(\tilde{\pi}_1 - \tilde{\pi}_2) \pm z_{\alpha/2} \sqrt{\frac{\tilde{\pi}_1(1-\tilde{\pi}_1)}{n_1 + 2} + \frac{\tilde{\pi}_2(1-\tilde{\pi}_2)}{n_2 + 2}},$$

with

$$\tilde{\pi}_1 = \frac{n_{11} + 1}{n_1 + 2} \qquad \text{and} \qquad \tilde{\pi}_2 = \frac{n_{21} + 1}{n_2 + 2}.$$

In this case, the Agresti-Caffo CI and the Wald CI are equal up to 4 decimal places.

Difference of proportions - In R

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

95% Wald CI:

```
prop.test(c(189, 104),      # vector of successes
          c(11034, 11037), # vector of group totals
          correct = FALSE  # no continuity correction
          )
```

Output:

```
---------------------------
95 percent confidence interval:
 0.004687751 0.010724297
---------------------------
```

Difference of proportions - In R

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

95% score CI (e.g., for small samples):

```
library(PropCIs)
diffscoreci(189, 11034, 104, 11037,
            conf.level = .95)
```

Output:

```
---------------------------
95 percent confidence interval:
 0.004716821 0.010788501
---------------------------
```

Ratio of proportions (relative risk)

Ratio of proportions (relative risk)

Instead of the difference $(\pi_1 - \pi_2)$, we can look at the ratio of the two proportions:

$$\boxed{\text{Relative risk} = \frac{\pi_1}{\pi_2}}$$

Motivation:
The same difference is more relevant when both proportions are near 0 or 1 than when they are close to .5.

Example:

- Case 1: $\pi_1 = .51$ and $\pi_2 = .50$.
- Case 2: $\pi_1 = .011$ and $\pi_2 = .001$.

In both cases, $\pi_1 - \pi_2 = .01$.

However:

- In Case 1, $\frac{\pi_1}{\pi_2} = 1.02$ and thus $(\pi_1 - \pi_2)$ is 2% of π_2.
- In Case 2, $\frac{\pi_1}{\pi_2} = 11$ and thus $(\pi_1 - \pi_2)$ is 1000% of π_2.

Ratio of proportions (relative risk)

$$\boxed{\text{Relative risk} = \frac{\pi_1}{\pi_2}}$$

The relative risk is:

- Always ≥ 0.
- Equal to 1 when $\pi_1 = \pi_2$ (i.e., when X and Y are independent).
- Larger than 1 when $\pi_1 > \pi_2$.
- Smaller than 1 when $\pi_1 < \pi_2$.

Ratio of proportions (relative risk)

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
			22071

$$\text{Sample relative risk} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{189/11034}{104/11037} = 1.82.$$

Thus, although $\hat{\pi}_1 - \hat{\pi}_2 = .0077$ is very small, we conclude that the proportion of MI is 82% higher for the placebo group in comparison to the aspirin group.

Ratio of proportions (relative risk)

The CI for the relative risk is complex.

Let's rely on *software* to compute it.

Ratio of proportions (relative risk) - In R

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

95% score CI for relative risk:

```
library(PropCIs)
riskscoreci(189, 11034, 104, 11037,
            conf.level = .95)
```

Output:

```
---------------------------
95 percent confidence interval:
 1.433904 2.304713
---------------------------
```

Ratio of proportions (relative risk) - Example

$$\boxed{95\% \text{ CI} = (1.434, 2.305)}$$

This interval is entirely above 1.

Thus, there is evidence against $\mathcal{H}_0 : \frac{\pi_1}{\pi_2} = 1$.

Conclusion:
We again conclude that *evidence suggests* that taking aspirin diminishes the risk of heart attack.

Exercise3-1

According to recent UN figures, the annual gun homicide rate is 62.4 per one million residents in the US and 1.3 per one million residents in Britain. Compare these two proportions of residents killed annually by guns using the (a) difference of proportions, (b) relative risk. Which measure is more useful for describing the strength of association? Why?

The odds ratio

The odds

The odds ratio is the ratio of two odds of the type

$$\text{odds} = \frac{\pi}{1-\pi},$$

where π is a probability of success.

For example, if $\pi = .8$ then the odds equal $\frac{.8}{1-.8} = 4$:

The probability of success is 4 times the probability of failure.
In other words, we expect to observe 4 successes for every one failure.

The success probability is a function of the odds:

$$\pi = \frac{\text{odds}}{\text{odds} + 1}.$$

Reverse-engineering the example above, we have that $\pi = \frac{4}{4+1} = .8$.

The odds and the odds ratio

	Y categories		
X categories	Y_1	Y_2	Total
X_1	n_{11}	n_{12}	n_1
X_2	n_{21}	n_{22}	n_2

We can compute the odds of success for each group (=row):

$$\text{odds}_1 = \frac{\pi_1}{1-\pi_1} \quad , \quad \text{odds}_2 = \frac{\pi_2}{1-\pi_2}.$$

The odds ratio is the ratio of these two odds:

$$\boxed{\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}}$$

This is different from the relative risk, $\frac{\pi_1}{\pi_2}$!

The odds ratio

$$\boxed{\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}}$$

Properties of the odds ratio:

- Always ≥ 0.
- $\theta = 1$ when X and Y are independent.
- $\theta > 1$ when $\pi_1 > \pi_2$.
- $\theta < 1$ when $\pi_1 < \pi_2$.
- $\frac{\text{odds}_1}{\text{odds}_2} = 1/\frac{\text{odds}_2}{\text{odds}_1}$.
  E.g., if the odds of success are 4 times higher in Group 1 than in Group 2, then they are .25 times as high in Group 2 than in Group 1.

Thus, the order of the groups (=rows) is immaterial.

The odds ratio

	Y categories		
X categories	Y_1	Y_2	Total
X_1	n_{11}	n_{12}	n_1
X_2	n_{21}	n_{22}	n_2

The *sample* odds ratio are given as follows:

$$\hat{\theta} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{\frac{n_{11}}{n_1}\Big/\frac{n_{12}}{n_1}}{\frac{n_{21}}{n_2}\Big/\frac{n_{22}}{n_2}} = \cdots = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

The odds ratio is therefore called the *cross-product ratio*.

The odds ratio

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

$$\hat{\theta} = \text{Sample odds ratio } = \frac{189 \times 10933}{10845 \times 104} = 1.83.$$

The estimated odds of MI for those taking placebo are 1.83 times the estimated odds for those taking aspirin.

The odds ratio

The sampling distribution of $\hat{\theta}$ is skewed to the right, especially for small sample sizes. Thus, normal approximations for $\hat{\theta}$ are unsuitable.

However, the sampling distribution of $\log\left(\hat{\theta}\right)$ is approx. normally distributed (for large samples):

$$\log\left(\hat{\theta}\right) \sim \mathcal{N}(\log(\theta), SE),$$

where

$$SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Under independence ($\pi_1 = \pi_2$), $\theta = 1$ and therefore $\log(\theta) = 0$.

The odds ratio – CI

$$\boxed{\log\left(\hat{\theta}\right) \sim \mathcal{N}(\log(\theta), SE)}$$

The large-sample $100(1-\alpha)\%$ Wald CI for $\log(\theta)$ is therefore given by

$$\log\left(\hat{\theta}\right) \pm z_{\alpha/2}(SE).$$

To get the $100(1-\alpha)\%$ Wald CI for θ we 'exponentiate' the interval above:

$$\exp\left\{\log\left(\hat{\theta}\right) \pm z_{\alpha/2}(SE)\right\}.$$

The odds ratio – CI

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

We saw before that $\hat{\theta} = \frac{189 \times 10933}{10845 \times 104} = 1.83$.

The 95% Wald CI for $\log(\theta)$:

$$\begin{aligned} CI_{log} &= \log\left(\hat{\theta}\right) \pm z_{\alpha/2}(SE) \\ &= \log(1.83) \pm 1.96 \times \sqrt{\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933}} \\ &= (0.365, 0.846). \end{aligned}$$

The 95% Wald CI for θ:

$$\begin{aligned} CI &= \exp\left\{\log\left(\hat{\theta}\right) \pm z_{\alpha/2}(SE)\right\} \\ &= \exp\left(0.365, 0.846\right) \\ &= (1.44, 2.33). \end{aligned}$$

The odds ratio – CI

Group	Myocardial infarction		Total
	Yes	No	
Placebo	189	10845	11034
Aspirin	104	10933	11037

$$CI = (1.44, 2.33)$$

Note that:

- $\hat{\theta} = 1.83$ is not the midpoint of the CI.
  This is because of the skewness of the sampling distribution of $\hat{\theta}$.
- There is evidence against $\pi_1 = \pi_2$, because 1 is excluded from the CI.

The odds ratio – CI

Wald's CI works for large samples.

For small samples, especially when π_1 and π_2 are close to 0 or 1, it is better to use the score CI.

Like before, we will only use software to compute the score CI.

The odds ratio – CI in R

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

95% Wald CI for odds ratio:

```
library(epitools)
oddsratio(c(189, 10845, 104, 10933),
          method  = "wald",
          correct = FALSE   # no continuity correction
         )
```

Output:

```
---------------------------

          odds ratio with 95% C.I.
Predictor  estimate    lower   upper
           1.832054 1.440042 2.33078
---------------------------
```

The odds ratio – CI in R

Group	Myocardial infarction Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

95% score CI for odds ratio:

```
library(PropCIs)
orscoreci(189, 11034, 104, 11037,
          conf.level = .95
         )
```

Output:

```
---------------------------
95 percent confidence interval:
 1.440802 2.329551
---------------------------
```

Odds ratio vs relative risk

$$\underbrace{\frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)}}_{\text{odds ratio}} = \underbrace{\frac{\hat{\pi}_1}{\hat{\pi}_2}}_{\text{relative risk}} \times \frac{1-\hat{\pi}_2}{1-\hat{\pi}_1}$$

When $\hat{\pi}_1$ and $\hat{\pi}_2$ are both close to zero then $\frac{1-\hat{\pi}_2}{1-\hat{\pi}_1} \simeq 1$ and therefore

odds ratio $\simeq$ relative risk.

Odds ratio vs relative risk

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\hat{\pi}_1 = \frac{189}{11034} = .0171$ and $\hat{\pi}_2 = \frac{104}{10933} = .0095$ are close to 0.

And indeed the odds ratio and the relative risk are close to each other:

$$\text{sample odds ratio} = \frac{189 \times 10933}{104 \times 10845} = 1.83.$$

$$\text{sample relative risk} = \frac{189/11034}{104/11037} = 1.82.$$

Thus, *exceptionally*, we can interpret odds ratio as relative risk:

> *The estimated probability of MI for the placebo group is 1.83 times the probability of MI for the aspirin group.*

Exercise 3-2

Consider the following two studies reported in the *New York Times*:
a. A British study reported that, of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer." Is 1.7 a relative risk?
b. A National Cancer Institute study about tamoxifen and breast cancer reported that the women taking the drug were 45% less likely to experience invasive breast cancer, compared to the women taking placebo. Find the relative risk for (i) those taking the drug compared to those taking placebo, (ii) those taking placebo compared to those taking the drug.

Exercise 3-3

For adults who sailed on the Titanic on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4.

a. What is wrong with the interpretation, "The probability of survival for females was 11.4 times that of males"? Give the correct interpretation.

b. The odds of survival for females equaled 2.9. For each gender, find the proportion who survived. Find the value of RR in the interpretation, "The probability of survival for females was RR times that for males."

Chi-squared tests of independence

Chi-squared tests of independence

Typically we wish to test these two hypotheses against each other:

$\mathcal{H}_0$: X and Y are independent.
$\mathcal{H}_1$: X and Y are dependent.

In order to perform significance testing, it's important to understand how counts are *expected* to be when $\mathcal{H}_0$ is true.

Chi-squared tests of independence – Expected counts

$$\boxed{\mathcal{H}_0 : X \text{ and } Y \text{ are independent}}$$

As we learned before, X and Y independent means that

$$\underbrace{P(X = i, Y = j)}_{\pi_{ij}} = \underbrace{P(X = i)}_{\pi_{i+}}\ \underbrace{P(Y = j)}_{\pi_{+j}},$$

for all $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

Thus, the expected count in the ij-th cell is given by

$$\mu_{ij} := n\pi_{ij} = n\pi_{i+}\pi_{+j}.$$

Chi-squared tests of independence – Expected counts

$$\boxed{\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}}$$

Estimation of the expected counts:

$$\begin{aligned}
\hat{\mu}_{ij} &= n\hat{\pi}_{i+}\hat{\pi}_{+j} \\
&= n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) \\
&= \frac{n_{i+}n_{+j}}{n} \\
&= \frac{\text{row total } \times \text{column total}}{\text{sample size}}
\end{aligned}$$

Chi-squared tests of independence – Example

Observed counts:

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

Expected counts under $\mathcal{H}_0$:

	Myocardial infarction		
Group	Yes	No	Total
Placebo	$\frac{11034 \times 293}{22071} = 146.48$	$\frac{11034 \times 21778}{22071} = 10887.52$	11034
Aspirin	$\frac{11037 \times 293}{22071} = 146.52$	$\frac{11037 \times 21778}{22071} = 10890.48$	11037
Total	293	21778	22071

Chi-squared tests of independence

$\mathcal{H}_0$: X and Y are independent.
$\mathcal{H}_1$: X and Y are dependent.

If $\mathcal{H}_0$ is true, then the *observed* $\{n_{ij}\}$ and the *expected* $\{\mu_{ij}\}$ counts should be similar.

Test statistics were developed to assess how dissimilar $\{n_{ij}\}$ and $\{\mu_{ij}\}$ are.

Pearson statistic and the chi-squared distribution

$\mathcal{H}_0$: X and Y are independent.
$\mathcal{H}_1$: X and Y are dependent.

The Pearson chi-squared statistic for testing $\mathcal{H}_0$ is

$$\boxed{X^2 = \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} = \sum_{i,j} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}}$$

As long as:

- The total sample size is not too small,

and

- No expected cell count μ_{ij} is smaller than (say) 5,

then

$$X^2 \underset{\text{under } \mathcal{H}_0}{\sim} \chi^2 \left(df = (r-1)(c-1)\right).$$

Pearson statistic and the chi-squared distribution

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2\left((r-1)(c-1)\right)$$

About X^2:

- It is always ≥ 0.
- It is exactly 0 if $\mathcal{H}_0$ is exactly true, i.e., if all $n_{ij} = \mu_{ij}$.
- The larger the differences between $\{n_{ij}\}$ and $\{\mu_{ij}\}$, the larger X^2.

Thus, the test's *critical region* is right-tailed.

Pearson statistic – Example

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189 (146.48)	10845 (10887.52)	11034
Aspirin	104 (146.52)	10933 (10890.48)	11037
Total	293	21778	22071

Obs (expected)

$$
\begin{aligned}
X^2 &= \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \\
&= \frac{(189 - 146.48)^2}{146.48} + \cdots + \frac{(10933 - 10890.48)^2}{10890.48} \\
&= 25.01.
\end{aligned}
$$

$$
p\text{-value} = \underbrace{P(X^2 > 25.01)}_{\chi^2(1)} < .001
$$

Conclusion: We reject the null hypothesis and conclude that X and Y are dependent.

Pearson statistic – In R

Pearson chi-squared test:

```
chisq.test(matrix(c(189, 10845, 104, 10933), ncol=2),
           correct = FALSE)
```

Output:

```
----------------------------
X-squared = 25.014, df = 1, p-value = 5.692e-07
----------------------------
```

Likelihood-ratio statistic

$\mathcal{H}_0$: X and Y are independent.

$\mathcal{H}_1$: X and Y are dependent.

The likelihood-ratio chi-squared statistic for testing $\mathcal{H}_0$ is

$$
\begin{aligned}
G^2 &= 2\log\left(\frac{\text{MLE under } \mathcal{H}_1}{\text{MLE under } \mathcal{H}_0}\right) \\
&= 2\sum_{i,j} n_{ij}\log\left(\frac{n_{ij}}{\mu_{ij}}\right) \\
&= 2\sum_{i,j} \text{observed} \times \log\left(\frac{\text{observed}}{\text{expected}}\right).
\end{aligned}
$$

Similarly to Pearson's X^2,

$$
G^2 \underset{\text{under } \mathcal{H}_0}{\sim} \chi^2\left(df = (r-1)(c-1)\right).
$$

Pearson statistic versus likelihood-ratio statistic

For large samples, they have similar values and follow the same asymptotic chi-squared distribution.

Likelihood-ratio statistic – Example

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189 (146.48)	10845 (10887.52)	11034
Aspirin	104 (146.52)	10933 (10890.48)	11037
Total	293	21778	22071

Obs (expected)

$$
\begin{aligned}
G^2 &= 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right) \\
&= 2\left[189 \times \log\left(\frac{189}{146.48}\right) + \cdots + 10933 \times \log\left(\frac{10933}{10890.48}\right)\right] \\
&= 25.37.
\end{aligned}
$$

$$
p\text{-value} = \underbrace{P(X^2 > 25.37)}_{\chi^2(1)} < .001
$$

Conclusion: We reject the null hypothesis and conclude that X and Y are dependent.

Likelihood-ratio statistic – In R

```r
MI.mat   <- matrix(c(189, 10845, 104, 10933), ncol = 2)
MI.chisq <- chisq.test(MI.mat, correct = FALSE)

with(MI.chisq, 2 * sum(observed * log(observed / expected)))
```

Output:

```
[1] 25.37196
```

Standardized residuals for cells in a contingency table

Residual $n_{ij} - \hat{\mu}_{ij}$, standardized assuming $\mathcal{H}_0$ holds, for the ij-th cell:

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - \hat{\pi}_{i+})(1 - \hat{\pi}_{+j})}}$$

For large samples,

$$r_{ij} \underset{\mathcal{H}_0}{\sim} \mathcal{N}(0, 1).$$

Rule-of-thumb:

- With few cells: $|r_{ij}| > 2$ indicates lack of fit of $\mathcal{H}_0$ for the ij-th cell.
- With many cells: $|r_{ij}| > 3$ indicates lack of fit of $\mathcal{H}_0$ for the ij-th cell.

Standardized residuals for cells in a contingency table

	Myocardial infarction		
Group	Yes	No	Total
Placebo	189 (146.48)	10845 (10887.52)	11034
Aspirin	104 (146.52)	10933 (10890.48)	11037
Total	293	21778	22071

Obs (expected)

$$r_{11} = \frac{189 - 146.48}{\sqrt{146.48\left(1 - \frac{11034}{22071}\right)\left(1 - \frac{293}{22071}\right)}} = 5.00$$

Similarly,

$$r_{12} = -5.00, \quad r_{21} = -5.00, \quad r_{22} = 5.00.$$

(Note: Residuals are symmetric columnwise.)

Conclusion:
For all cells, the residuals indicate lack of fit of $\mathcal{H}_0$.

Standardized residuals – In R

Standardized residuals:

```
chisq.test(matrix(c(189, 10845, 104, 10933), ncol=2),
           correct = FALSE)$stdres
```

Output:

```
---------------------------
          [,1]      [,2]
[1,]  5.001388 -5.001388
[2,] -5.001388  5.001388
---------------------------
```

Exercise 3-4

The below table shows data from a 2002 General Society Survey cross-classifying a person's perceived happiness with their family income. The table displays the observed and expected cell counts and the standardized residuals for testing independence.

Table: Data with observed and estimated frequencies, and standardized residuals

	Happiness		
Income	Not Too Happy	Pretty Happy	Very Happy
Above average	21 (35.8) −2.973	159 (166.1) −0.947	110 (88.1) 3.144
Average	53 (79.7) −4.403	372 (370.0) 0.224	221 (196.4) 2.907
Below average	94 (52.5) 7.368	249 (244.0) 0.595	83 (129.5) −5.907

(a) Show how to obtain the expected cell count of 35.8 for the first cell.

(b) For testing independence, $X^2 = 73.4$. Report the df value and the P-value, and interpret (use $\alpha = 5\%$).

(c) Interpret the standardized residuals in the corner cells, having observed counts 21 and 83.

(d) Interpret the standardized residuals in the corner cells, having observed counts 110 and 94.

Exercise 3-5

The Pearson chi-squared statistic formula presented on slide 57 has an alternative form: $X^2 = n\sum(\hat{\pi}_{ij} - \hat{\pi}_{i+}\hat{\pi}_{+j})^2/\hat{\pi}_{i+}\hat{\pi}_{+j}$. Explain why, for fixed $\hat{\pi}_{ij}$ values, X^2 becomes large when n is sufficiently large, regardless of whether the association is practically important.

Note: In particular, what the result above shows is that the chi-squared test merely indicates the degree of evidence against the null hypothesis of independence. The chi-squared test does not describe the *strength* of the association between both variables.

In the next lecture

We are going to skip Chapter 3.4.5 and 3.5 from the Chapter 3 of the textbook.