

✓ ~ CDA I → ST3241
CDA II → ST4...

Categorical Data Analysis

Cate. var

Topic 4. summarize 1 quant var:
_____ association between 2 var:

W & H

W & gender

Topic 6

gender & lung cancer

- 1 Introduction
- 2 Summaries of a Single Categorical Variable
- 3 Two Categorical Variables: 2×2 Tables
 - Contingency Tables
 - Prospective Versus Retrospective Studies
 - Chi-squared (χ^2) Test for 2×2 Contingency Tables
- 4 Chi-squared (χ^2) Test for $r \times c$ Tables
 - Two Nominal Variables
 - Table With Ordinal Variable

1 Introduction

2 Summaries of a Single Categorical Variable

3 Two Categorical Variables: 2×2 Tables

- Contingency Tables
- Prospective Versus Retrospective Studies
- Chi-squared (χ^2) Test for 2×2 Contingency Tables

4 Chi-squared (χ^2) Test for $r \times c$ Tables

- Two Nominal Variables
- Table With Ordinal Variable

Categorical Variables

Gender: M & F ; $\begin{matrix} 0 = 1 \\ 1 = 0 \end{matrix}$ 2 cate: M & F M, I, C, O
 $M_1 = F_0 = 1$ C Non C

- A variable is called categorical variable if each observation belongs to one of a set of categories.
- Examples of categorical variables are gender, religion, race, type of residence.
- Distinguishing between quantitative and categorical variables: simply asking if there is a meaningful distance between any two points in the data. If such a distance is meaningful then you have quantitative data.
For instance, it makes sense to compute the difference in systolic blood pressure between subjects but it does not make sense to consider the mathematical operation ("smoker" - "non-smoker").
- It is important to identify which type of data you have (quantitative or categorical), as it affects the exploration techniques that you can apply.
- A categorical variable is **ordinal** if the observations can be ordered, but do not have specific quantitative values. *gender*
- A categorical variable is **nominal** if the observations can be classified into categories, but the categories have no specific ordering.
- The purpose in this chapter is to identify any association between two categorical variables and how to do that with software.

1 Introduction

2 Summaries of a Single Categorical Variable

3 Two Categorical Variables: 2×2 Tables

- Contingency Tables
- Prospective Versus Retrospective Studies
- Chi-squared (χ^2) Test for 2×2 Contingency Tables

4 Chi-squared (χ^2) Test for $r \times c$ Tables

- Two Nominal Variables
- Table With Ordinal Variable

Summaries of a Single Categorical Variable

gender : 80 M
20 F

- For a single categorical variable, we can use frequency table (which also can produce the proportion or percentage of categories) as numerical summaries. The category with the highest frequency is the modal category.

hist is for quant var

- A common graphical to display a categorical variable is bar plot.

→ for cate var

Summaries of a Single Categorical Variable in R

```
> data <- read.csv("C:/Data/bats.csv")
```

```
> count = table(data$type)
```

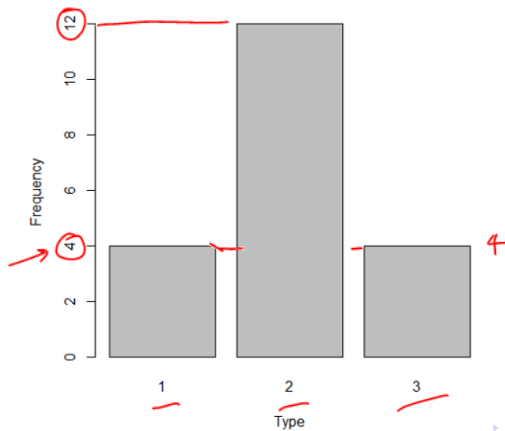
```
> count # frequency table
```

① ② ③

④ 12 ④

Model cate. is type 2.

```
> barplot(count)
```



Summaries of a Single Categorical Variable in Python (1)

mass type er...

```
import pandas as pd
data = pd.read_csv(r"C:/Data/bats.csv")
tab = pd.crosstab(index=data["type"], columns="count")
print(tab)
```

col_0	<u>count</u>
<u>type</u>	
1	4
2	12
3	4

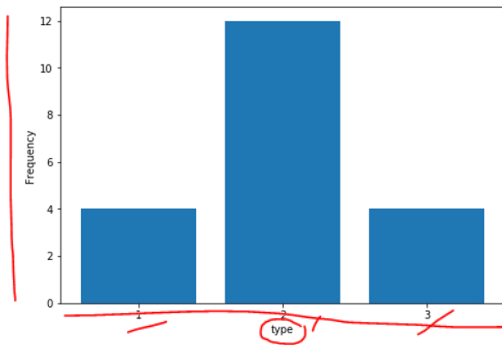
→ 1
→ 2
→ 3

data. type

frequency

Summaries of a Single Categorical Variable in Python (2)

```
import matplotlib.pyplot as plt  
fig = plt.figure()  
ax = fig.add_axes([0,0,1,1])  
type = ['1', '2', '3']  
counts = [4, 12, 4]  
ax.bar(type, counts)  
plt.xlabel('type')  
plt.ylabel('Frequency')  
plt.show()
```

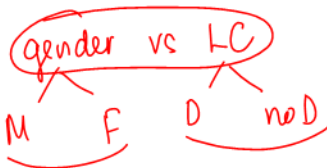


1 Introduction

2 Summaries of a Single Categorical Variable

4 x 5 ..

2 x 3 ..

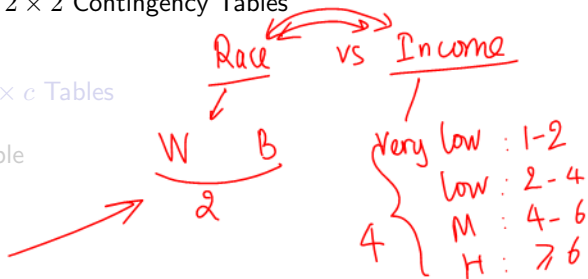


3 Two Categorical Variables: 2 x 2 Tables

- Contingency Tables
- Prospective Versus Retrospective Studies
- Chi-squared (χ^2) Test for 2 x 2 Contingency Tables

4 Chi-squared (χ^2) Test for $r \times c$ Tables

- Two Nominal Variables
- Table With Ordinal Variable



Two Categorical Variables

There are different methods to explore the association of two categorical variables

- Contingency table (compare proportions; obtain odds ratio)
- Chi-square test (χ^2)
- Charts

A Simple Example

Example (Chest Pain And Gender)

- Suppose that 1073 patients at NUH were sampled, for a study where the onset of severe chest pain in patients at high risk for cardiovascular disease (CVD) is recorded for each subject.
- The 1073 patients were queried on two aspects:
 - Have they experienced the onset of severe chest pain in the preceding 6 months? (yes/no)
 - Gender? (male/female)

Data

	Chest Pain	No Chest Pain	Total
Male	46	474	520
Female	37	516	553
Total	83	990	1073

- 1 Introduction
- 2 Summaries of a Single Categorical Variable
- 3 Two Categorical Variables: 2×2 Tables
 - Contingency Tables
 - Prospective Versus Retrospective Studies
 - Chi-squared (χ^2) Test for 2×2 Contingency Tables
- 4 Chi-squared (χ^2) Test for $r \times c$ Tables
 - Two Nominal Variables
 - Table With Ordinal Variable

Contingency Tables

- The two categorical variables are gender and presence/absence of chest pain. We can compute **conditional proportions** in the table for the Chest Pain example. CP

	Chest Pain	No Chest Pain	Total
Male	8.8%	91.2%	100%
Female	6.7%	93.3%	100%

gender is
explanatory

$$Pr(CP | Male) = 8.8\%$$

$$Pr(CP | Female) = \frac{37}{553} = 6.7\%$$

- Some questions to answer: Is the probability of having chest pain in male is the same as the probability of having chest pain in female?
Are the two variables independent (or associated)?
Can we infer the causality (gender causing different probability of having chest pain)?
- The table above is called a **contingency table**. It is important to identify which variable is the response variable and which one is the explanatory variable, so that the conditional proportion is calculated correctly for making inference.

$$\frac{46}{520}$$

$$\frac{8.8\%}{8\%}$$

Contingency Tables: Comparing Proportions

$$\Pr(CP | \text{Male}) \longleftrightarrow \Pr(CP | \text{Female})$$

- Let's assume that we have response Y with 2 outcomes in the columns (success and failure) and explanatory X has 2 levels/categories in rows.

$$\begin{aligned} \pi_1 &= \Pr(CP | \text{Male}) \text{ in pop} \leftarrow p_1 \text{ in sample} = 8.8\% \\ \pi_2 &= \Pr(CP | \text{Female}) \text{ in pop} \leftarrow p_2 \text{ in sample} = 6.7\% \end{aligned}$$

- In row 1 and row 2, let π_1 , π_2 denote the probabilities of a success, respectively.

Let p_1 and p_2 denote the sample proportion of successes in row 1 and row 2.

- 1073
- The sample difference $p_1 - p_2$ is used to estimate the real difference $\pi_1 - \pi_2$. If this difference is significant, we can infer the association between X and Y .

- Relative risk: The ratio p_1/p_2 is used to estimate π_1/π_2 . If this RR is significantly different from 1, we also can infer the association between X and Y .

$$\frac{8.8}{6.7}$$

Contingency Tables: Odds Ratio

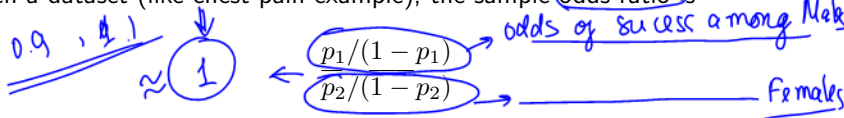
2x2

- Another term that can help to explore the association in a 2-way table is odds ratio.
- For a probability of success π , the odds of success is defined as $odds = \pi / (1 - \pi)$.
- In the 2-way contingency table here, the odds ratio (OR) is the ratio of two odds:

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

whereas the relative risk is a ratio of two probabilities.

- Given a dataset (like chest pain example), the sample odds ratio is



Odds ratio can be any non-negative number.

The Properties of the Odds Ratio

- When X and Y are independent, in the population, we have $\pi_1 = \pi_2$, so $\theta = 1$. This value is a baseline for comparison.
- The further values of θ from 1 in a given direction, the stronger association it represents.
- If the order of the rows or the order of the columns is reversed (but not both), the new value of θ is the inverse of the original value.
This ordering is usually arbitrary, so whether we get $\theta = 4$ or $\theta = 0.25$ is simply a matter of how we label the rows and columns.
- The odds ratio does not change when the table orientation reverses so that the rows become the columns and the columns become the rows. This also means that the OR takes the same value when it is defined using the conditional distribution of X given Y as it does when defined using the distribution of Y given X . That is, it treats the variables symmetrically. This is the big difference between OR and RR or OR and the difference of proportions.

Confidence Intervals for Odds Ratio

A 2×2 table for 2 variables x and Y has 4 cell counts $n_{11}, n_{12}, n_{21}, n_{22}$, the sample OR is

$$\hat{\theta} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} \quad \rightarrow \quad \hat{\theta} = 1.353$$

- 100%(1 - α) confidence interval for the real odds ratio θ is formed by

$$\exp \left\{ \log \hat{\theta} \pm z_{\alpha/2} \times \text{ASE}(\log \hat{\theta}) \right\}$$

Asymptotic SE of $\log \hat{\theta}$

CI of $\log \theta$

where

Agresti

$$\text{ASE}(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

- If we want to get a 95% CI for the population OR then we'll use $\alpha = 0.05$, and $z_{\alpha/2} = 1.96$. from $N(0,1)$

- If the CI contains 1, that means the population OR might be 1, hence two variables X and Y might be independent!

CI for θ is $\rightarrow (1.2, 1.8)$; $(0.8, 3)$

Chest Pain Example (1)

- We are interested in knowing if these two conditional proportions in the **population** are similar. If the two conditional proportions are (very) different from one another, we say that there is an association between gender and chest pain. If they are similar, we say that there is no association, or that the two variables are independent.
- $p_1 = 8.8\%$ is a point estimate of $P(\text{chest pain}|\text{male})$ in population. Similarly, $p_2 = 6.7\%$ is a point estimate of $P(\text{chest pain}|\text{female})$.
- The proportion difference from the sample is: 2.1% (not very big difference); sample RR is 1.3 (not very different from 1).
- The sample odds ratio is $\hat{\theta} = 1.353$. This can be interpreted as: in the given sample, the odds of having chest pain among males is 1.353 times the odds of having chest pain among females.

Chest Pain Example (1)

- The 95% CI for population OR is

$$\exp\left\{\log 1.353 \pm 1.96 \times 0.23\right\} = \underline{(0.86, 2.12)}$$

- The 95% CI contains 1, hence we can infer that gender and chest pain might not be associated.

- 1 Introduction
- 2 Summaries of a Single Categorical Variable
- 3 Two Categorical Variables: 2×2 Tables
 - Contingency Tables
 - Prospective Versus Retrospective Studies
 - Chi-squared (χ^2) Test for 2×2 Contingency Tables
- 4 Chi-squared (χ^2) Test for $r \times c$ Tables
 - Two Nominal Variables
 - Table With Ordinal Variable

Suitable Methods to Assess the Association

	Positive Outcome (Disease)	Negative Outcome (No disease)
Exposure	a	b
No exposure	c	d

- Consider a 2×2 contingency table as above, where X is the exposure variable with 2 outcomes (exposure and no exposure) and variable Y with 2 outcomes (yes and no). There are at least 3 measures of association available:

- the difference between two conditional proportions:

$$\pi_1 \leftarrow P_1$$

$$\pi_2 \leftarrow P_2$$

$$\hat{p}_1 - \hat{p}_2 = \frac{a}{a+b} - \frac{c}{c+d}$$

$$P_1 - P_2 \leftarrow 1^{st}$$

$$P_1/P_2 \leftarrow 2^{nd}$$

$$\leftarrow 3^{rd}$$

- the relative risk: \hat{p}_1/\hat{p}_2 .

- the odds ratio: $\hat{\theta} = \frac{a/b}{c/d} = \frac{ad}{bc}$

$$\frac{P_1/(1-P_1)}{P_2/(1-P_2)} = \frac{ad}{bc}$$

- However, not all the sample are suitable (or valid) to use all these 3 measures. Only the prospective studies can use all these measure to quantify the association between X and Y , whereas the retrospective studies can use odds ratio only.

Prospective Study

Smoking $\begin{cases} \text{yes} \\ \text{no} \end{cases}$

LC

1. Sample subjects randomly from a population.
2. Either randomly assign the exposure variable to the subjects, or record their exposure variable status.
3. Follow the subjects over time to see if they develop the disease.

This is the main advantage:

- Can obtain valid estimate of \hat{p}_1 and \hat{p}_2 from the 2×2 table. Hence, all the three measures of association are valid for this kind of study.

Retrospective Study

$$\begin{array}{l} \Pr(D|Ex) \\ \Pr(D|noEx) \end{array}$$

D
↙
↘
NoD

$$\begin{array}{l} \Pr(Ex|D) \\ \Pr(Ex|noD) \end{array}$$

- 1 Sample a group of cases (people with the disease).
- 2 Sample a group of controls (people without the disease).
- 3 Check each subject to see if they were exposed or not.

This is also known as a *case-control* study. These are the advantages:

- Cheap
- Quick
- Fewer subjects involved, especially if disease is rare.

However, the huge disadvantage is that we cannot obtain valid estimate of \hat{p}_1 and \hat{p}_2 from the 2×2 table, since we need the estimate of $\Pr(\text{disease}|\text{exposure})$ whereas from the retrospective study we got $\Pr(\text{exposure}|\text{disease})$ instead.

Retrospective or Prospective?

Experimental → prospective

observational

Ro

- Sample 5000 OC users and 5000 non-OC users, and follow them for 15 years to see if they develop any form of myocardial infarction.

Re

- Identify and sample breast cancer cases in mothers at a hospital. From the same hospital, identify and obtain a sample similarly aged mothers, but who do not have breast cancer. Now check for their age at which they had their first child (record as greater than 30 or not).

Contingency Tables in R

proportion, RR, OR, CI for OR

	CP	No CP
Male	46	474
Fe	37	516
		<u>0.5</u>

```
> chest.pain<-matrix(c(46,474,37,516), ncol=2, byrow=2)
> dimnames(chest.pain)<-list(Gender=c("Male", "Female"),
+                             CP=c("Yes", "No"))
> ##2-sample test for equality of proportions without
> ##continuity correction:
> test<-prop.test(chest.pain,correct=FALSE)
> RR<-(test$estimate[1])/(test$estimate[2])
> odds<-test$estimate/(1- test$estimate)
> OR<-odds[1]/odds[2]
```



Contingency Tables in R

Building a function to OR and CI of OR:

```
> OR<-function(x, pad.zeros = FALSE, conf.level=0.95){  
+   if(pad.zeros){if(any(x==0)) {x<-x+0.5}}  
+   theta<-x[1,1]*x[2,2]/(x[2,1]*x[1,2])  
+   ASE<-sqrt(sum(1/x))  
+   CI<-exp(log(theta) +c(-1,1)*qnorm(0.5*(1+conf.level))*ASE)  
+   list(estimator=theta, ASE=ASE,conf.interval=CI,  
+        conf.level=conf.level) }  
> OR(chest.pain)
```

\$estimator

```
[1] 1.353404
```

\$ASE

```
[1] 0.2298126
```

\$conf.interval

```
[1] 0.8626023 2.1234612
```

\$conf.level

```
[1] 0.95
```

Contingency Tables in Python

```
tab = [ df['Yes']/(df['Yes'] + df['No']), df['No']/(df['Yes'] + df['No']) ]  
tab = np.asmatrix([tab[0],tab[1]])  
tab = np.transpose(tab)  
print('Conditional Probabilities ', '\n', tab)
```

```
Conditional Probabilities  
[[0.08846154 0.91153846]  
 [0.06690778 0.93309222]]
```

```
data = {'Yes': [46,37], 'No': [474,516]}  
  
df = pd.DataFrame(data, columns=['Yes', 'No'])  
prob = df['Yes']/(df['Yes'] + df['No'])  
print('RR = ', prob[0]/prob[1]) #this is Relative Risk  
  
odds = prob/(1-prob) #the odds of 'Yes'  
print('OR = ', odds[0]/odds[1]) ### this is odds ratio
```

```
RR = 1.3221413721413722  
OR = 1.3534040369483409
```

1 Introduction

2 Summaries of a Single Categorical Variable

3 Two Categorical Variables: 2×2 Tables

- Contingency Tables
- Prospective Versus Retrospective Studies
- Chi-squared (χ^2) Test for 2×2 Contingency Tables

4 Chi-squared (χ^2) Test for $r \times c$ Tables

- Two Nominal Variables
- Table With Ordinal Variable

Gender & CP

- indpd + large samples size
- indpd + small size
- dpndnt samples

Independence and Dependence of Two Categorical Variables



$$\begin{aligned} &\rightarrow \Pr(CP | M) \\ &\rightarrow \Pr(CP | Fe) \end{aligned}$$

Arrows point from the text "Gender" to the conditions "M" and "Fe" in the probability expressions above.

Definition (Independence and Dependence (Association))

- Two categorical variables are **independent** if the population conditional distributions for one of them are identical at each category of the other.
- The variables are **dependent**, or associated, if the conditional distributions are not identical.

We now learn about a hypothesis test (χ^2 test) for association between two categorical variables.

Chi-square Tests: Hypotheses

- The test we are about to learn has the following hypotheses:

$$\begin{aligned} H_0 &: \text{The two variables are independent} \Rightarrow \theta = 1 \\ H_1 &: \text{The two variables are dependent} \Rightarrow \theta \neq 1 \end{aligned}$$

- For those who are unfamiliar with procedures of performing a hypothesis test: we'll need to find "evidence" that against H_0 from the given data. If the evidence is strong enough, we'll tend to reject H_0 and conclude that the two variables might be dependent.
- In order to find the evidence against H_0 , we'll assume X and Y are independent as stated under H_0 and calculate what are the cell counts should be, called expected counts. The more these expected counts are similar as the observed counts the weaker the evidence (against H_0); if these expected counts are very different from the observed counts then we got strong evidence against H_0 .
- For a particular cell, the **expected count** is

$$\text{Expected count} = \frac{\text{Row total} \times \text{Column total}}{\text{Total sample size}}$$

Chi-square Tests: Test Statistic

- In short, test statistic is the evidence that the data provide. It will summaries how far the observed counts are from the expected counts.

- A larger value of the test statistic will give stronger evidence against the null hypothesis.

- The formula χ^2 test statistic (with continuity correction) is:

$\chi^2 \sim \chi^2_{distn}$ \rightarrow $\chi^2 = \sum \frac{(|\text{observed count} - \text{expected count}| - 0.5)^2}{\text{expected count}}$ $\sim \chi^2_{df}$

Handwritten notes: $\sum \frac{(E - O)^2}{E}$ (with arrow to 'Raw'), $(|E - O| - 0.5)^2$ (with arrow to 'continuity correction'), and $\chi^2 \sim \chi^2_{df}$ (with arrow to 'distribution').

Note that there are variations on this formula in other books/documents.

- Expected counts are not necessarily integers. If all the expected counts are larger than 5, it's suitable to use chi-square test to test the independence. **If there is at least one expected count lesser than 5, we say sample is of small size, and there is another test for this situation.**

2×2 table.

Chi-square Tests: p-value and Conclusion

the smaller p-value, the strong evidence against H_0 .

- p-value helps to quantifies how strong the evidence against H_0 is.
- p-value is calculated from χ^2 distribution with degree of freedom 1, which is the the area in the right of the test statistic value.
- The smaller p-value the stronger the evidence against H_0 is.

- 0.01*
- 0.01 - 0.05*
- Conclusion: If p-value is small (< 0.05) we can say: data provide strong evidence against H_0 . If p-value is moderately small (between 0.05 and 0.2) we say data provide evidence against H_0 but not strong. If p-value is large (> 0.2) we say data do not provide enough evidence against H_0 .

- 0.05*
- Given a pre-specified significance level α (the common α is 0.05), if p-value $< \alpha$ we'll reject H_0 ; Otherwise we do not reject H_0 .
- 0.06 < 0.1*

Chi-square Tests: Chest Pain Example

We'll perform a chi-square test of independence between gender and chest pain at significance level $\alpha = 0.05$.

- The expected counts are:

H_0 :
 H_a :

	Chest Pain	No Chest Pain	Total
Male	40.2	479.8	520
Female	42.8	510.2	553
Total	83	990	1073

All the expected counts are larger than 5.

- Test statistic: $1.456 = \frac{(40.2 - 46)^2}{40.2} + \frac{(479.8 - 474)^2}{479.8} + \dots + \dots$
- p-value: 0.228 follows χ^2_1 -distribution.
- Conclusion: Data do not provide enough evidence against H_0 ; At $\alpha = 0.05$, we do not reject H_0 and conclude that gender and chest pain might be independent.

H_0 : indpd

Chi-square Tests in R and Python

- In R:

```
> chisq.test(chest.pain)
```

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

data: chest.pain

X-squared = 1.4555, df = 1, p-value = 0.2276

- In Python:

```
import scipy.stats as scst  
obs = np.array([[46,474], [37,516]])  
print(obs)  
scst.chi2_contingency(obs, correction = True)
```

```
[[ 46 474]  
 [ 37 516]]
```

```
(1.4555294041803708, → Test statistic  
 0.2276427809700174, → p-value  
 1, → Degree of freedom  
 array([[ 40.22367195, 479.77632805],  
        [42.77632805, 510.22367195]]))
```

Small Sample Sizes

- If any of the expected counts in the 2×2 table is lesser than 5, we should use Fisher exact test.
- This test uses the same hypotheses as the chi-square test, the procedures to perform the test is similar, however the test statistic is obtained differently and the way p-value is calculated is also different.
- Fisher exact test should be used for the example below

Example (Claritin and Nervousness)

- Claritin is a drug for treating allergies. However, it has a side effect of inducing nervousness in patients.
- From a sample of 450 subjects, 188 of them were randomly assigned to take Claritin, and the remaining were assigned to take the placebo. Data were recorded:

Handwritten notes: $P(N | \text{Claritin})$, $P(N | \text{placebo})$, P_1 , P_2 , $OR > 1$, \downarrow

	Nervous	Not Nervous	Total
Claritin	4	184	188
Placebo	2	260	262

Handwritten notes: OR, RR , $\neq 1$, $H_1: \theta > 1$

Fisher Exact Test in R

- 50% of the expected counts (2 expected counts) are smaller than 5, hence using chi-square test is not suitable.

- Fisher exact test gives p-value of 0.2412.

```
> claritin <- matrix(c(4, 184, 2, 260), ncol=2, byrow=2)
> fisher.test(claritin, alternative = "two.sided")
```

Chisq Fisher's Exact Test for Count Data

data: claritin

p-value = 0.2412

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.399349 31.473382

sample estimates:

odds ratio

2.819568

H_0 : 2 variables are indep

$\Rightarrow \theta = 1$

$H_1: \theta \neq 1$

$\left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}$ 1-sided
 $\neq 1$ 2 sided

$\theta \neq 1$

95% CI for θ : (0.4; 31.5)

Fisher Exact Test in Python

```
import scipy.stats as scst  
claritin = np.array([[4,184], [2,260]])  
print(claritin)  
scst.fisher_exact(claritin, alternative='two-sided')
```

```
[[ 4 184]  
 [ 2 260]]
```

```
(2.8260869565217392, 0.24118420183181116)
```

p-value

Dependent samples in 2×2 Tables

- It may happen that samples in a 2×2 table are dependent. Example below is a case.
- There are 50 students taking a statistical course. The lecturer gave a test on R at the beginning of the course. There were 26 students who passed and 24 of them failed. After taking the course which can help students improve their ability working with R, another test was given. This time, 42 students passed and 8 of them failed. Does taking this course help students to improve their ability working with R?

data \Rightarrow

		After		Total
		Pass	Fail	
Before	Pass	25 ^a	1 ^b	26
	Fail	17 ^c	7 ^d	24
	Total	42	8	50

matched
paired
dependent

- The samples for Before and for After are the collected from the same set of 50 students. Hence, the two samples are dependent. Thus, the chi-square test or Fisher exact test should not be used for this case.

Dependent samples in 2×2 Tables: McNemar's Test

- McNemar's test should be used in this case. If the table is having the cell counts denoted as a, b, c, d then the idea is: to see if the statistical course effectively improve students' ability in using R, we need to check the count of b, c . 24 students who failed in Before, now 17 of them move to Pass under After ($c = 17$), whereas 26 who passed under Before now 1 of them move to Fail under After ($b = 1$). The moving of these numbers 17 and 1 is just a randomness or is it because of the dependence?

- The null hypothesis: the course is of NO help.

H_0 : the results of Before and After are independent, or: the course has no effect

- The test statistic (used for the case when sample is large enough):

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

Raw, no correction
→ for large table

or when sample has a small cell count, we can use the test statistic ~~with~~ with correction:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

$$\frac{(|E - 0.5|)^2}{E}$$

- The test statistic above follows a χ^2 distribution.

1 Introduction

2 Summaries of a Single Categorical Variable

3 Two Categorical Variables: 2×2 Tables

- Contingency Tables
- Prospective Versus Retrospective Studies
- Chi-squared (χ^2) Test for 2×2 Contingency Tables

2×2 } 3 cases

$r \times c$ where at least r or $c \geq 2$

3×2 2×6

3×5

4 Chi-squared (χ^2) Test for $r \times c$ Tables

- Two Nominal Variables
- Table With Ordinal Variable

Categorical Variables With More Than 2 Categories

- We have considered the situation of two categorical variables where each one has only two outcomes (2×2 table).
- It's very common that we want to check the association between two nominal variables where one of them or both have more than 2 outcomes. The more special case is one or both variables are ordinal.
- Consider data given in the table below where both variables are nominal

Example (Gender Gap in Political Affiliation)

Gender	Democrat	Independent	Republican	Total
Females	762	327	468	1557
Males	484	239	477	1200
Total	1246	566	945	2757

Table: Cross Classification of Party Identification by Gender

- 1 Introduction
- 2 Summaries of a Single Categorical Variable
- 3 Two Categorical Variables: 2×2 Tables
 - Contingency Tables
 - Prospective Versus Retrospective Studies
 - Chi-squared (χ^2) Test for 2×2 Contingency Tables
- 4 Chi-squared (χ^2) Test for $r \times c$ Tables
 - Two Nominal Variables
 - Table With Ordinal Variable

Chi-squared (χ^2) Test for $r \times c$ Tables

- The χ^2 test can be extended to tables larger than 2 by 2.
- In general suppose that we have r rows and c columns that define two categorical random variables.
- Expected value in each cell is computed exactly the same way as for the 2×2 table.
- The only difference is that the χ^2 distribution to use is the $\chi^2_{(r-1)(c-1)}$ distribution (the degree of freedom is $(r-1)(c-1)$).

```
> political <- matrix(c(762, 327, 468, 484, 239, 477), ncol=3, byrow=2)
```

```
> chisq.test(political)
```

Pearson's Chi-squared test

data: political

X-squared = 30.07, df = 2, p-value = 2.954e-07

provide strong evidence
against $H_0 \Rightarrow$ it suggests
the association between
Gender & PA.

Standardized Residuals

- Test statistics and p-value describe the evidence against H_0 . A cell-by-cell comparison of observed and estimated expected frequencies is necessary to help us to understand better the nature of evidence.
- Define standardized residual (or **adjusted residual**)

$$r_{ij} = \frac{n_{ij} - \mu_{ij}}{\sqrt{\mu_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

Standard error of $(n_{ij} - \mu_{ij})$

where n_{ij} is the observed count in row i and column j (cell ij); μ_{ij} is the expected count for cell ij under H_0 ; p_{i+} is the marginal probability of row i and p_{+j} is the marginal probability of column j .

$\sqrt{\mu_{ij}(1 - p_{i+})(1 - p_{+j})}$ is the estimated standard error of $(n_{ij} - \hat{\mu}_{ij})$ under H_0 .

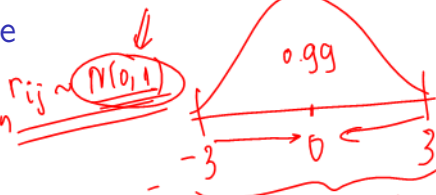
- The residuals r_{ij} can be derived by the output of chi-square test.
- When H_0 is true, each r_{ij} has a large-sample standard normal distribution. If $|r_{ij}|$ in a cell exceeds 2 then it indicates lack of fit of H_0 in that cell.

Standardized Residuals: Example

```
> chisq.test(political)$stdres
```

Democrat *Indpd* *Republican*

	[,1]	[,2]	[,3]
Female	4.502054	0.6994517	-5.315946
Male	-4.502054	-0.6994517	5.315946



More males like Rep than H0 expects.

- Keep in mind that H_0 states the independence between gender and political affiliation.
- Large positive residuals for female Democrats (4.5) and male Republicans (5.3), thus there were *more* female Democrats and male Republicans than H_0 predicts.
- Large negative residuals for female Republicans (-5.3) and male Democrats (-4.5), thus there were **fewer** female Republicans and male Democrats than the H_0 predicts.
- One could consider the 2×2 table of Democrat and Republican to obtain the sample odds ratio and have further interpretation.

Some Comments about Chi-Squared Tests

- Pearson's χ^2 test only indicate the degree of evidence for an association, but they usually can not answer other questions about dataset. Better to study the nature of the association, rather than relying solely on these tests.

We'll use \rightarrow stronger condition \rightarrow $< 25\%$ of cells have $EC < 5$.

- χ^2 test is not always applicable, since they require large samples (so that the sampling distribution of χ^2 can be closer to chi-squared distribution).

The approximation is poor when $n/(IJ) < 5$.

n = total sample size

I = no of rows; J = no of columns

weaker condition

\rightarrow $\geq 75\%$ of cells have $EC \geq 5$

- Another test that is equivalent to the chi-square test for $r \times c$ where $r \geq 2$ and $c \geq 2$ (and even better in some situation) is the likelihood ratio test.

Students can find out more about how to conduct this test by self-studying.

- χ^2 does not depend on the order in which the rows and columns are listed. Thus they ignore some information when there is ordinal variable.

Income X Gender

- 1 Introduction
- 2 Summaries of a Single Categorical Variable
- 3 Two Categorical Variables: 2×2 Tables
 - Contingency Tables
 - Prospective Versus Retrospective Studies
 - Chi-squared (χ^2) Test for 2×2 Contingency Tables
- 4 Chi-squared (χ^2) Test for $r \times c$ Tables
 - Two Nominal Variables
 - Table With Ordinal Variable

Testing Independence for Ordinal Data

at least 1 variable
is ordinal.

X = nominal \rightarrow Gender $\begin{matrix} M \leftarrow \\ F \end{matrix}$

- It is quite common to assume that as the levels of X increases, responses on Y tend to increase or to decrease toward higher levels of X.

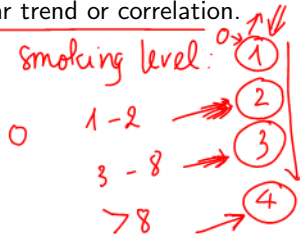
- To detect a trend association, most simple and common analysis assigns scores to categories and measures the degree of linear trend or correlation.

total I rows & J columns

- Let u_1, u_2, \dots, u_I denote scores for the rows.

- Let v_1, v_2, \dots, v_J denote scores for the columns.

Smoking level:



- The scores have the same ordering as the category level, they should reflect distances between categories (greater distances between categories regarded as further apart).

Linear-by-Linear Association Test

or 1 nominal and 1 ordinal

- A test for the association of 2 ordinal variables is linear-by-linear test. Its null hypothesis is: two variables are independent; the alternative hypothesis is: two variables are dependent.

- The test statistic is calculated by

$$M^2 = (n-1)r^2$$

$M^2 \sim \chi^2$
 $(\Rightarrow) M \sim N(0,1)$

sample size

correlation, below

$u_1 \dots u_4$

where r is sample correlation between X and Y , $\bar{u} = \sum_i u_i p_{i+}$ is the sample mean of row scores, $\bar{v} = \sum_j v_j p_{+j}$ is the sample mean of the column scores.

$$r = \frac{\sum_{i,j} (u_i - \bar{u})(v_j - \bar{v}) p_{ij}}{\sqrt{[\sum_i (u_i - \bar{u})^2 p_{i+}] [\sum_j (v_j - \bar{v})^2 p_{+j}]}}$$

Note that: $p_{ij} = n_{ij}/n$; $p_{i+} = n_{i+}/n$; $p_{+j} = n_{+j}/n$.

- For large samples, test statistic M^2 has approximately a chi-squared distribution with 1 degree of freedom.

for r x c table

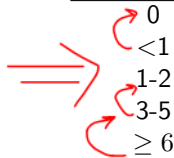
Linear-by-Linear Association Test: Example (1)

Example (Alcohol Use and Infant Malformation) 5×2

<u>Alcohol</u> Consumption	<u>Malformation</u> \leftarrow Response		
	Absent	Present	Total
$\rightarrow 0 \leftarrow$	17,066	48	17,114
$\rightarrow < 1$	14,464	38	14,502
$\rightarrow 1-2$	788	5	793
$\rightarrow 3-5$	126	1	127
$\rightarrow \geq 6$	37	1	38
Total	32481	93	32574

Table: Infant Malformation and Mother's Alcohol Consumption

Linear-by-Linear Association Test: Example (2)



Alcohol Consumption	Malformation			Percentage Present
	Absent	Present	Total	
0	17,066	48	17,114	0.28
<1	14,464	38	14,502	0.26
1-2	788	5	793	0.63
3-5	126	1	127	0.79
≥ 6	37	1	38	2.63

Table: Infant Malformation and Mother's Alcohol Consumption

From the table, it seems when the Alcohol Consumption increases, the Percentage Present and the Standardized Residual increase. This suggest that there may be a linear trend.

Linear-by-Linear Association Test: Example (3)

- Assign scores to alcohol consumption: $v_1 = 0, v_2 = 0.5, v_3 = 1.5, v_4 = 4, v_5 = 7$ and $u_1 = 0, u_2 = 1$. \rightarrow *mal*
- we have $r = 0.014, n = 32574, M = 2.56$.
- Consider H_1 : the association between two variables is positive (one sided test), then we get p-value = 0.00519. The two sided p-value (for H_1 : two variables are dependent) is 0.01.
- It suggests strong evidence of a linear trend for infant malformation with alcohol consumption of mothers that the more alcohol used by the mother during pregnancy the larger probability of infant malformation.
- Note that, the choice of the scores may affect the result. If we had taken $v = (1; 2; 3; 4; 5)$, then $M^2 = 1.83$ and two-sided p-value = 0.18 gives a much weaker conclusion.
- It is usually better to use one's own judgment by selecting scores that reflect distances between categories.

Linear-by-Linear Association Test in Python

- How the test is conducted in R, you can check the file `Topic6_Rcode.R`
- Homework: write the code to conduct linear-by-linear association test for the Infant Malformation sample in Python.