Measures of Association and Pearson Chisquare Test of Independence

Cheng Peng

STA200 Statistics II

Contents

1 Introduction

We have discussed the **goodness-of-fit test for categorical distributions**. As previously stated, a goodness-of-fit test for a categorical distribution is a special comparison between an **observed frequency distribution** (also called the empirical distribution) and a **hypothetical distribution**. In other words, the χ^2 **goodness-of-fit test** assesses a specific relationship between two distributions: the **empirical distribution** and the **hypothetical distribution**.

This module focuses on assessing a **more general relationship between two categorical variables** (or, equivalently, two categorical populations). From this perspective, there are **two extreme relationships** between two categorical variables:

1. **The two distributions are identical**: We have already discussed the χ^2 goodness-of-fit test for assessing this relationship.
2. **The two variables are independent** (i.e., no association): We will discuss the χ^2 **test of independence** at the end of this module.

In the next few sections, we will:

- Introduce **contingency tables** to summarize the joint distributions of two categorical variables.
- Define **measures of association** for two binary categorical variables.
- Discuss the well-known χ^2 **test of independence** for two categorical variables (including binary categorical variables).

2 Epidemiologic Study Designs

In descriptive statistics, we introduce a frequency table and a bar chart to characterize the distribution. We also briefly introduced chi-square tests of independence of two categorical variables. We will review chi-square tests and related tests and perform data analysis using R in the subsequent module 4.

To help you gain a better understanding of various chi-square tests to assess the association between two categorical variables. For illustration, we focus on the case of two binary categorical variables. Some of the concepts in the following can be generalized to the case of categorical variables with multiple categories.

Three study designs are fundamental in epidemiology and clinical research in which two categorical variables are involved: **outcome variable** (such as disease status) and **exposure variable** (such as smoking status). Each has distinct strengths, weaknesses, and applications. Below is a structured comparison.

2.1 Cohort Study Design

A cohort study is a type of epidemiological study in which a group of people with a **common characteristic** is **followed over time** to find how many reach a certain health **outcome** of interest (disease, condition, event, death, or a change in health status or behavior). A

Cohort studies compare an exposed group of individuals to an unexposed (or less exposed) group of individuals to determine if the **outcome** of interest is associated with **exposure**. That is, cohort studies focus on the relationship between the outcome and exposure variables.

Data Collection in Cohort Studies is based on stratified sampling. That is, the entire population is divided into subpopulations according to the values of the outcome variable or the exposure variable. Sub-random samples are then taken from each subpopulation, respectively.

- If the population is stratified by the outcome variable - the status of lung cancer, the two stratified samples are taken from the cancer population and the cancer-free population.
- If the population is stratified by the exposure variable - the status of smoking, the two stratified samples are taken from the smoking population and the non-smoking population.

The combined sample is called a stratified sample. There are two types of cohort studies: **prospective** and **retrospective** (or historical) cohorts.

- **Prospective studies** follow a cohort into the future for a health outcome. This means the exposure variable splits the population into the exposed population and the unexposed population. The two subsamples will be taken from the exposed population and the unexposed population, respectively, and then followed up for some time to observe the outcome.
- **retrospective studies** trace the cohort **back** *in time* for exposure information after the outcome has occurred. The retrospective cohort study is also called a **case-control study**.

Caution:

2.2 Cross-sectional Study Design

A **cross-sectional study design** is a type of observational research method that analyzes data from a population, or a representative subset, **at a single point in time**. It is commonly used in epidemiology, public health, social sciences, and market research.

Most of the statistical models were developed based on cross-sectional data. For example, a researcher surveys 1,000 adults to assess the prevalence of hypertension and collects information about age, BMI, and smoking status **at the time** of the survey.

2.3 Randomized Controlled Trials

A **Randomized Controlled Trial (RCT)** is a **prospective experimental study design** considered the gold standard for evaluating the effectiveness of interventions (e.g., drugs, treatments, policies).

- **Randomization**: Participants are randomly assigned to either an intervention group or a control group (e.g., placebo or standard treatment).
- **Control**: A comparison group is used to measure the effect of the intervention.
- **Blinding (optional but common)**: Reduces bias. Can be:
  - Single-blind (participants unaware)
  - Double-blind (participants and researchers unaware)
- **Prospective**: Follows participants forward in time after assignment.

Example: A clinical trial randomly assigns 200 patients with high blood pressure to receive either a new antihypertensive drug or a placebo. Blood pressure is monitored over 6 months to evaluate the drug's effectiveness.

The following short YouTube video summarizes the above major study designs (https://www.youtube.com/watch?v=_fJKfulldBM).

3 Contingency Tables and Measures of Association

This section is **descriptive** in nature. We begin by introducing **contingency tables** to describe the joint distribution of two **general categorical variables**, including binary categorical variables. The measures of association introduced in this section, however, are defined **exclusively** for **two binary categorical variables**.

3.1 Contingency Tables

Contingency tables (also called **cross-tabulations** or **crosstabs**) are fundamental tools in statistics for analyzing relationships between categorical variables. They organize data into rows and columns to display frequency distributions, enabling researchers to identify patterns, test hypotheses, and measure associations.

Structure of Contingency Tables

The general structure of a contingency table is depicted in the following:

	Outcome (Yes)	Outcome (No)	Total
Exposure (yes)	a	b	a + b
Exposure (No)	c	d	c + c
Total	a + c	b +d	a + b + c + d

We can see that **a basic contingency table** is an $r \times c$ matrix where:
+ **Rows** (r): Represent categories of one variable.
+ **Columns** (c): Represent categories of another variable.
+ **Cells**: Contain frequency counts for each variable combination.

The above contingency table is essentially a two-way (or bivariate) frequency table. Similar to what we did in introductory statistics (MAT121/125 at WCU), we can turn the above **raw (ordinary) frequency table** into the corresponding **relative bivariate frequency table** in the following form, where T = a + b + c + d.

	Outcome (Yes)	Outcome (No)	Total
Exposure (yes)	a/T	b/T	(a + b)/T
Exposure (No)	c/T	d/T	(c + c)/T
Total	(a + c)/T	(b +d)/T	T= a + b + c + d

Important note on the layout of contingency table: *Column names MUST be the distinct values of OUTCOME variable and row names MUST be the distinct values of EXPOSURE variable!!!*

The following YouTube video explains the above contingency table with an example (https://www.youtube.com/watch?v=W95BgQCp_rQ).

3.2 Risk Measures of Association

For a single categorical variable, we focus on the distribution using frequency tables and charts. In the case of two categorical variables, we focus primarily on the association between them. The basic analytic logic is to assess whether the association between the two categorical variables exists; if it does, we need to define numerical measures to measure the strength of the association.

We have introduced different study designs for data collection. A dataset collected using **Different study designs** contains **different amounts of information**. This means that when analyzing a contingency table and defining measures of association, we need to know the study design associated with the contingency table.

Here are commonly used risk measures based on the following general 2-by-2 contingency table.

	Disease (+)	Disease (-)	Total
Exposed (+)	a	b	a+b
Unexposed (-)	c	d	c+d

	Disease (+)	**Disease (-)**	**Total**
Total	a+c	b+d	N=a+b+c+d

1. Absolute Risk Measures

- Risk in Exposed (Attack Rate) is the probability of disease in the exposed group, which is defined by

$$AR_{\text{exp}} = \frac{a}{a+b}.$$

- Risk in Unexposed is the probability of disease in the unexposed group, which is defined by

$$AR_{\text{unexp}} = \frac{c}{c+d}.$$

2. Relative Risk Measures

- Risk Ratio (RR)

$$RR = \frac{AR_{exp}}{AR_{unexp}} = \frac{a/(a+b)}{c/(c+d)}$$

- Odds Ratio (OR)

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc},$$

where a/b is the odds of disease and c/d the odds of disease-free.

- **Interpretation**
- RR (OR) = 1: No association
- RR (OR)> 1: Increased risk with exposure
- RR (OR)< 1: Protective effect
- **When to use**:

Design	Primary Measure	Formula
Case-Control	Odds Ratio (OR)	$OR = ad/bc$
Cohort	Relative Risk (RR)	$RR = \frac{a/(a+b)}{c/(c+d)}$
Cross-Sectional	Prevalence Ratio	$\frac{\text{Prevalence}_1}{\text{Prevalence}_2}$

4 Chi-square Test of Independence

Let X and Y be two categorical variables with k and m categories, respectively. Their relationship between X and Y is characterized by their joint distribution (table). For simplicity, we use the following two special categories to explain the ideas of statistical testing of independence.

4.1 Independence of Two Categorical Variables

We use the following example to illustrate **independence** and **dependence** between two categorical variables.

Example 5. *Joint probabilities and contingency tables.* Let X = political preference (Democrat vs Republican) and Y =gender (Male and Female). Let's **assume** their joint distribution to be of the following *contingency table*.

	Democrat	Republican	Row total
Male	$p_{11} = 0.3$	$p_{12} = 0.2$	$p_{1*} = 0.50$
Female	$p_{21} = 0.3$	$p_{22} = 0.3$	$p_{2*} = 0.50$
Column Total	$p_{*1} = 0.50$	$p_{*2} = 0.50$	$p_{**} = 1.00$

Column marginal probabilities

Row marginal probabilities

The cell numbers are joint probabilities. For example, $p_{12} = 0.2 = 20\%$ says 20% of the *study population* are male republicans. The row and column totals represent the percentage of males/females and democrats/republicans in the study population. Any observed data table is **governed** by the above joint distribution table.

Definition Two categorical variables are **independent** if and only if their joint probabilities are equal to the product of their corresponding marginal probabilities.

With this definition, we can see that X and Y with joint distribution specified in the above table (in Example 5) are NOT independent since $p_{11} = 0.3 \neq 0.5 \times 0.5 = 0.25$.

Example 6. We consider two variables X = preference of hair color (Blonde and Brunette) and Y = gender (Male and Female). Assume the joint distribution of the two variables is given by

	Male	Female	**total**
Blonde	0.18	0.27	**0.45**
Brunette	0.22	0.33	**0.55**
total	**0.40**	**0.60**	**1.00**

Based on the definition of independence. The preference for hair color is independent of gender. Since **all** joint probabilities are equal to the product of their corresponding marginal probabilities.

$0.45 \times 0.40 = 0.18, 0.45 \times 0.60 = 0.27, 0.55 \times 0.40 = 0.22$, and $0.55 \times 0.60 = 0.33$.

4.2 Expected Table Under Independence Assumption (H_0)

We construct the **expected table** under the **null hypothesis of independence** and the **observed contingency table**. For ease of interpretation, we use an example to illustrate the steps for obtaining the expected table.

Example 7. Consider the potential dependence between the attendance (good vs poor) and course grade (pass vs fail). We take 50 students from a population and obtain the following observed table.

	Pass	Fail	total
Good	25	2	27
Poor	8	15	23
total	33	17	50

Question: Is attendance independent of class performance?

$$Ho: \ attendance \ is \ \ independent \ \ of \ the \ performance$$

versus

$$Ha: \ attendance \ is \ \ dependent \ of \ the \ performance$$

To obtain the expected table, we follow the next few steps.

1. Estimate the marginal probabilities

	Pass	Fail	Marginal Probability
Good			0.54
Poor			0.46
Marginal Probability	0.66	0.34	1.00

where marginal probabilities are calculated by Pr(Good) = 27/50 = 0.54, Pr(Poor) = 23/50 = 0.46, Pr(Pass) = 33/50 = 0.66, Pr(Fail) = 17/50 = 0.34.

2. Estimate the joint probability under the null hypothesis of independence

	Pass	Fail	Marginal Probability
Good	0.3564	0.1836	0.54
Poor	0.3036	0.1564	0.46
Marginal Probability	0.66	0.34	1.00

where joint probabilities under the independence assumption (H_0) are calculated by taking the product of the corresponding marginal probabilities. For example, $0.54 \times 0.66 = 0.3564$.

3. Calculate the Expected Table

The expected frequencies are calculated in the following table (with detailed steps).

	Pass	Fail	Row Total
Good	0.3564 x 50 =**17.82**	0.1836 x 50 = **9.18**	27
Poor	0.3036 x 50 =**15.18**	0.1564 x 50 =**7.82**	23
Column Total	33	17	Sample size = 50

Remark. For categorical variables with **more than two categories**, the expected table can be found using ** the same 3 steps** as those used in the above example.

4.3 Formulation of Chi-squared Test of Independence

The test statistic used to test the independence of two categorical variables is the same as that used in the goodness-of-fit test. That is the standardized "distance" between the observed and the expected table (under H_0).

Assume that the two categorical variables have k and m categories respectively, then the resulting test statistic has a chi-square distribution with $(k-1) \times (m-1)$ degrees of freedom.

Example 8. [Continuation of **Example 7**]. Test whether **attendance** and **class performance**.

Solution: We have found the expected table under H_0 in **Example 7**, we put the observed and expected tables in the following.

Observed Table

	Pass	Fail	total
Good	**25**	**2**	27
Poor	**8**	**15**	23
total	33	17	50

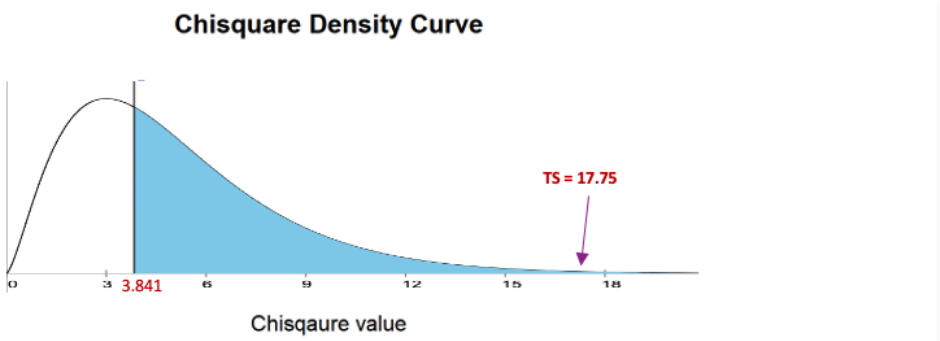
Expected Table

	Pass	Fail	total
Good	**17.82**	**9.18**	27
Poor	**15.18**	**7.82**	23
total	33	17	50

The test statistic is given by

$$TS = \frac{(25-18.82)^2}{17.82} + \frac{(2-9.18)^2}{9.18} + \frac{(8-15.18)^2}{15.18} + \frac{(15-7.82)^2}{7.82} = 17.75$$

The test statistic has a chi-square distribution with $(2-1)\times(2-1) = 1$ degrees of freedom. The critical value at the significance level of 0.05 is found in the following figure.



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.01
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.6
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.2
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.3
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143	13.2
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.0

Since the test statistic is inside the rejection region, we reject the null hypothesis that attendance and class performance are independent.

Example 9. Do some college majors require more studying than others? The National Survey of Student Engagement asked a number of college freshmen what their major was and how many hours per week they spent studying, on average. A sample of 1000 of these students was chosen, and the numbers of students in each category are tabulated in the following two-way contingency table.

Hours Studying Per Week	Major				Total
	Humanities	Social Science	Business	Engineering	
0–10	68	106	131	40	345
11–20	119	103	127	81	430
More Than 20	70	52	51	52	225
Total	257	261	309	173	1000

Solution: The null and alternative hypotheses are given by

```
Ho: studying time is INDEPENDENT of majors
                versus
Ha: studying time is DEPENDENT on majors.
```

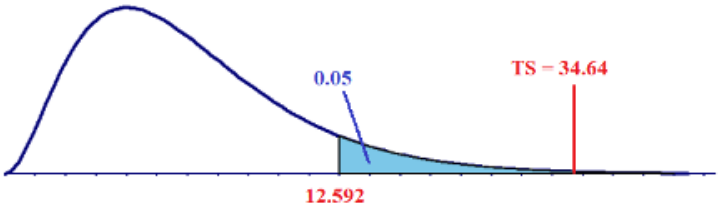
Under the null hypothesis, we obtained the expected table using the same steps as in Example 7 in the following.

	Humanities	Soc Sci	Business	Engineering	Total
0-10	88.7	90.0	106.6	59.7	345
11-20	110.5	112.2	132.87	74.4	430
>20	57.8	58.7	69.5	38.9	225
total	457	261	309	173	1000

The test statistic is given by

$$TS = \frac{(68-88.7)^2}{88.7} + \frac{(106-90.0)^2}{90.0} + \cdots + \frac{(52-38.9)^2}{38.9} \approx 34.64$$

The critical value and rejection region based on a significance level of 0.05 are given by



d.f	0.9950	0.9900	0.9750	0.9500	0.9000	0.1000	0.0500	0.0250	0.010
1	0.0000	0.0002	0.0010	0.0039	0.0158	2.705	3.841	5.024	6.63
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.605	5.992	7.378	9.21
3	0.0717	0.1148	0.2158	0.3518	0.5844	6.251	7.815	9.348	11.34
4	0.2070	0.2971	0.4844	0.7107	1.0636	7.779	9.488	11.143	13.27
5	0.4117	0.5543	0.8312	1.1455	1.6103	9.236	11.070	12.832	15.08
6	0.6757	0.8721	1.2373	1.6354	2.2041	10.645	12.592	14.449	16.81
7	0.9893	1.2390	1.6899	2.1673	2.8331	12.017	14.067	16.013	18.47

Conclusion: Since the test statistic is inside the rejection region, we reject the null hypothesis and conclude that the studying time is dependent on the majors.

To conclude this section, watch the following YouTube video for another manually worked-out example of the chi-squared test of independence.

4.4 Implementing χ^2 Independence Test in R

The R function `chisq.test()` in package `{stats}` takes either a contingency table (aggregated data) or a variable directly from a data set.

Based on Contingency Table

To use a contingency table, we first load the contingency table in R. Next, I

Hours Studying Per Week	Major				Total
	Humanities	Social Science	Business	Engineering	
0–10	68	106	131	40	345
11–20	119	103	127	81	430
More Than 20	70	52	51	52	225
Total	257	261	309	173	1000

The following R code defines a matrix to store the above contingency table.

```
# First step is to define a vector to store the cells in the above contingency table
cell.vec <- c(68, 106, 131, 40, 119, 103, 127, 81, 70, 52, 51, 52) # read values by row
## populate the table cells using the above vector by row using the R function matrix()
## matrix() requires the number of rows and the way of populating the table cell:
## byrow = TRUE (read in data by row), byrow = FALSE (read in data by column)
cont.table <- matrix(cell.vec, nrow = 3, byrow = TRUE)
## add row names and column names
rownames(cont.table) <- c("0-10", "11-20", "More Than 20")     # add row names
colnames(cont.table) <- c("Humanities", "Social-Science", "Business", "Engineering")  # add column name
cont.table
```

```
##              Humanities Social-Science Business Engineering
## 0-10                 68            106      131          40
## 11-20               119            103      127          81
## More Than 20         70             52       51          52
```

Once the R contingency table is defined, we can simply call the R function `chisq.test()` in the following code.

```
chisq.test(cont.table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont.table
## X-squared = 34.638, df = 6, p-value = 5.065e-06
```

This gives the same results as those obtained from manual calculation.

Based on Two Columns of Categorical Variables in the Dataset

Consider two variables, **outlook** and **sport.decision**, in the following artificial data set. The objective is to see whether **outlook** affects **Sport.decision**

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

To implement the χ^2 test of independence between **outlook** and **Sport.decision**, we define the two variables first and then call `chisq.test()`:

```
outlook <- c("Sunny", "Sunny", "Overcast", "Rain", "Rain", "Rain", "Overcast", "Sunny", "Sunny", "Rain"
sport.decision <-c("No", "No", "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes
## We can use the R function table() to cross-tabulate the above two variables to make a contingency ta
contingency <- table(outlook, sport.decision)
```

Next, we call the R function `chisq.test()` with both raw variables and the aggregated contingency table, respectively.

```
chisq.test(outlook, sport.decision)
```

```
##
##  Pearson's Chi-squared test
##
## data:  outlook and sport.decision
## X-squared = 3.5467, df = 2, p-value = 0.1698
```

```
chisq.test(contingency)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency
## X-squared = 3.5467, df = 2, p-value = 0.1698
```

The above two function calls yield the same results.

5 Practice Exercises Using R

1. **Political Affiliation and Opinion**: The following table, based on the sample, will be used to explore the relationship between Party Affiliation and Opinion on Tax Reform. Find the expected counts for all of the cells.

	favor	indifferent	opposed	total
democrat	138	83	64	285
republican	64	67	84	215
total	202	150	148	500

2. **Tire Quality**: The operations manager of a company that manufactures tires wants to determine whether there are any differences in the quality of work among the three daily shifts. She randomly

selects 496 tires and carefully inspects them. Each tire is either classified as perfect, satisfactory, or defective, and the shift that produced it is also recorded. The two categorical variables of interest are the shift and condition of the tire produced. The data can be summarized by the accompanying two-way table. Does the data provide sufficient evidence at the 5% significance level to infer that there are differences in quality among the three shifts?

	Perfect	Satisfactory	Defective	Total
Shift 1	106	124	1	231
Shift 2	67	85	1	153
Shift 3	37	72	3	112
Total	210	281	5	496

3. **Condiment preference and gender**: A food services manager for a baseball park wants to know if there is a relationship between gender (male or female) and the preferred condiment on a hot dog. The following table summarizes the results. Test the hypothesis with a significance level of 10%.

		Condiment			
		Ketchup	Mustard	Relish	Total
Gender	Male	15	23	10	48
	Female	25	19	8	52
	Total	40	42	18	100