Introduction to Applied Statistics
STAT 5005
Categorical Data (Chapter 10)

Jingyu Sun

Fall 2021

Introduction

Inferences about a Population Proportion π

Inferences about the Difference between Two Population Proportions, $\pi_1 - \pi_2$

Inferences about Several Proportions: Chi-Square Goodness-of-Fit Test

Contingency Tables: Test for Independence

Odds and Odds Ratios

Introduction

Introduction

- Up to this point, we have been concerned primarily with sample data measured on a quantitative scale
- In some situations, levels of the variable of interest are identified by name or rank only
- We are interested in the number of observations occurring at each level of the variable
- Data obtained from these types of variables are called **categorical** or **count data**
- Example: item coming off an assembly line may be classified into one of three quality classes: acceptable, repairable, or reject
- In this chapter, we will examine specific inferences that can be made from experiments involving categorical data

Inferences about a Population Proportion π

Inferences about a Population Proportion π

- Assume we have a binomial experiment. π is the probability of success. Then the probability distribution for y, the number of successes in n identical, is

$$P(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

- The point estimate of the binomial parameter π is $\hat{\pi} = y/n$
- When $n\pi \geq 5$ and $n(1-\pi) \geq 5$, the sampling distribution of $\hat{\pi}$ is approximated by a normal distribution with a mean and a standard error as given here

$$\mu_{\hat{\pi}} = \pi \text{ and } \sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

- A $100(1-\alpha)\%$ confidence interval for π is

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

- Example: A new genetic treatment of 870 patients with a particular type of cancer resulted in 330 patients surviving at least 5 years after treatment. Estimate the proportion of all patients with the specified type of cancer who would survive at least 5 years after being administered this treatment. Use a 90% confidence interval

- The sample size required for a $100(1-\alpha)\%$ confidence interval for π of the form $\hat{\pi} \pm E$ is

$$n = \frac{z_{\alpha/2}^2 \pi(1-\pi)}{E^2}$$

- Notes:
  - The specified confidence interval width is $W = 2E$
  - Since π is unknown, either substitute an educated guess or use $\pi = 0.5$
- Example: A new PC operating system is being developed. The designer wants to determine how many programs to randomly sample in order to estimate the proportion of Microsoft Windows-compatible programs that would perform adequately using the new operating system.
- The designer wants the estimator to be within 0.03 of the true proportion using a 95% confidence interval as the estimator

Statistical Test About a Population Proportion π

- Null hypothesis $H_0 : \pi = \pi_0$
- Three possible types of alternative hypotheses:
  1. $H_a : \pi > \pi_0$
  2. $H_a : \pi < \pi_0$
  3. $H_a : \pi \neq \pi_0$
- The test statistic is $z = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
- For a specified level of significance α, reject H_0 if (match corresponding case)
  1. $z > z_\alpha$
  2. $z < -z_\alpha$
  3. $|z| > z_{\alpha/2}$
- For valid inferences, n must satisfy $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$

Example

- Binge drinking is defined as consuming five or more drinks in a row three or more times in a two-week period
- An extensive survey of colleges students reported that 44% of U.S. college students engaged in binge drinking during the two weeks before the survey
- A service fraternity of a university conducted a survey of 2,500 undergraduates attending their university and found that 1,200 of the 2,500 students had engaged in binge drinking
- Is there sufficient evidence to indicate that the percentage of students engaging in binge drinking at the university is greater than the percentage found in the national survey? Use a significance level of 5%

Inferences about the Difference between Two Population Proportions, $\pi_1 - \pi_2$

- Many practical problems involve the comparison of two proportions
- Example: Compare smokers and non-smokers with respect to their opinions concerning imposing a federal tax to help pay for health care reform
- We assume that independent random samples are drawn from two binomial populations with unknown parameters designated by π_1 and π_2
- If y_1 (resp. y_2) successes are observed for the random sample of size n_1 (resp. n_2) from population 1 (resp. population 2), then the point estimate of π_1 (resp. π_2) is $\hat{\pi}_1 = y_1/n_1$ (resp. $\hat{\pi}_2 = y_2/n_2$)
- The sampling distribution for $\hat{\pi}_1 - \hat{\pi}_2$ can be approximated by a normal distribution with mean and standard error given by

$$\mu_{\hat{\pi}_1 - \hat{\pi}_2} = \pi_1 - \pi_2 \text{ and } \sigma_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

- A $100(1-\alpha)\%$ confidence interval for $\pi_1 - \pi_2$ is

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

- Example: A company test-markets a new product in two regions A and B
- In region A, the company spends a roughly equal dollar amount on a balanced mix of television, radio, newspaper, and magazine ads. The company's advertising in region B is based almost entirely on television commercials.
- Two months after the ad campaign begins, the company conducts surveys to determine consumer awareness of the product
  - 527 people interviewed in region A, 413 aware of the product
  - 608 people interviewed in region B, 392 aware of the product
- Calculate a 95% confidence interval for the regional difference in the proportion of all consumers who are aware of the product

Statistical Test for the Difference between Two Population Proportions

- Null hypothesis $H_0 : \pi_1 = \pi_2$
- Three possible types of alternative hypotheses:
  1. $H_a : \pi_1 - \pi_2 > 0$
  2. $H_a : \pi_1 - \pi_2 < 0$
  3. $H_a : \pi_1 - \pi_2 \neq 0$
- The test statistic is $z = \dfrac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}}$
- For a specified level of significance α, reject H_0 if (match corresponding case)
  1. $z > z_\alpha$
  2. $z < -z_\alpha$
  3. $|z| > z_{\alpha/2}$
- For valid inferences, $n_1\hat{\pi}_1, n_1(1-\hat{\pi}_1), n_2\hat{\pi}_2$, and $n_2(1-\hat{\pi}_2)$ must be all at least 5

Example (10.8)

An educational researcher designs a study to compare the effectiveness of teaching English to non-English-speaking people by a computer software program and by the traditional classroom system. The researcher randomly assigns 125 students from a class of 300 to instruction using the computer. The remaining 175 students are instructed using the traditional method.

```
Exam.Result <-  matrix(c(94,113,31,62,125,175),nrow=3,
                       byrow = TRUE)
colnames(Exam.Result) <-  c("Computer Instruction",
                            "Traditional Instruction")
rownames(Exam.Result) <- c("Pass", "Fail", "Total")
Exam.Result
```

```
##       Computer Instruction Traditional Instruction
## Pass                    94                      113
## Fail                    31                       62
## Total                  125                      175
```

Does instruction using the computer software program appear to

Fisher's Exact Test

- When at least one of the quantities $n_1\hat{\pi}_1, n_1(1-\hat{\pi}_1), n_2\hat{\pi}_2$, or $n_2(1-\hat{\pi}_2)$ is less than 5, the previous test procedure is invalid and the **Fisher Exact Test** should be used
- We need to develop the exact probability distribution for the cell counts in all 2×2 tables having the same row and column totals $n_1, n_2, m, N-m$

	Outcome		
Population	Success	Failure	Total
1	x	?	n_1
2	?	?	n_2
Total	m	$N-m$	N

- Under $H_0 : \pi_1 = \pi_2$, what is the probability of observing a particular value for x?

- Under $H_0 : \pi_1 = \pi_2$, the probability of observing a particular value for x is

$$P(x) = \frac{\binom{n_1}{x}\binom{n_2}{m-x}}{\binom{N}{m}}$$

- To test the difference in the two population proportions, the p-value of the test is the sum of these probabilities for outcomes at least as supportive of the alternative hypothesis as the observed table
- In R, $P(x)$ can be computed with the command below

```
dhyper(x,n1,n2,m)
```

Example

- A clinical trial is conducted to compare two drug therapies for leukemia: P and PV. Twenty-one patients were assigned to drug P and forty-two patients to drug PV

	Outcome		
Population	Success	Failure	Total
PV	38	4	42
P	14	7	21
Total	52	11	63

- Is there significant evidence that the proportion of patients obtaining a successful outcome is higher for drug PV than for drug P?

- The p-value is the sum of the probabilities for all tables having 38 or more successes:

$$
\begin{aligned}
P(x \geq 38) &= P(38) + P(39) + P(40) + P(41) + P(42) \\
&= \frac{\binom{42}{38}\binom{21}{14}}{\binom{63}{52}} + \frac{\binom{42}{39}\binom{21}{13}}{\binom{63}{52}} + \frac{\binom{42}{40}\binom{21}{12}}{\binom{63}{52}} + \frac{\binom{42}{41}\binom{21}{11}}{\binom{63}{52}} + \frac{\binom{42}{42}\binom{21}{10}}{\binom{63}{52}} \\
&= \texttt{sum(dhyper(38:42,42,21,52))} \\
&= 0.025365
\end{aligned}
$$

- Reject H_0 for any level of significance $\alpha \geq 0.026$

Inferences about Several Proportions: Chi-Square Goodness-of-Fit Test

The Multinomial Experiment

- A multinomial experiment has the characteristics listed here

1. The experiment consists of n identical trials.
2. Each trial results in one of k outcomes.
3. The probability that a single trial will result in outcome i is π_i for $i = 1, 2, \ldots, k$, and remains constant from trial to trial. (Note: $\sum_{i=1}^{k} \pi_i = 1$).
4. The trials are independent.
5. We are interested in n_i, the number of trials resulting in outcome i. (Note: $\sum_{i=1}^{k} n_i = n$)

- The probability distribution for the number of observations resulting in each of the k outcomes, called the multinomial distribution, is given by the formula

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k}$$

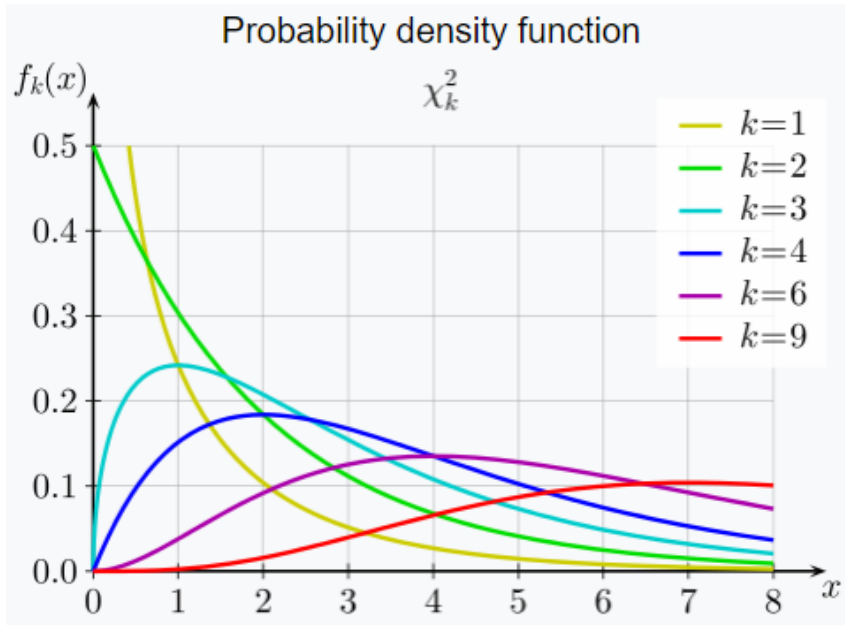
Inferences about Several Proportions: Chi-Square Goodness-of-Fit Test

- We want to test whether the sample data agree with the hypothesized values for the mutinomial probabilities $\pi_1, \ldots, \pi_k$
- The **expected number of outcomes** of type i in n trials is $E_i = n\pi_i$. The expected numbers $E_1, \ldots, E_k$ are also called **expected cell counts**
- The observed numbers $n_1, \ldots, n_k$ are called **observed cell counts**

- The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i}$$

- The distribution of the quantity χ^2 can be approximated by a **chi-square** distribution provided that the expected cell counts E_i are fairly large



- A chi-square distribution is specified by one parameter, called the number of degrees of freedom

Chi-Square Goodness-of-Fit Test

- Null hypothesis: $\pi_i = \pi_{i0}$ for categories $i = 1, \ldots, k$
- Alternative hypothesis: At least one of the cell probabilities differs from the hypothesized value
- Test statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - E_i)^2}{E_i}$
- For a specified significance level α, reject H_0 if χ^2 exceeds the critical value χ^2_α with $k - 1$ degrees of freedom
- In R, χ^2_α =`qchisq(1-alpha,k-1)`

Validity of the Chi-Square Distribution Approximation

- The approximation of the sampling distribution of the chi-square goodness-of-fit test statistic by a chi-square distribution improves as the sample size n becomes larger
- Cochran (1954) indicates that the approximation should be adequate if no E_i is less than 1 and no more than 20% of the E_is are less than 5

Example

- A laboratory is comparing a test drug to a standard drug preparation that is useful in the maintenance of patients suffering from high blood pressure
- The lab classifies the responses to therapy for a large patient group into one of four response categories (see table below)

Category	Percentage
1: Marked decrease in blood pressure	50
2: Moderate decrease in blood pressure	25
3: Slight decrease in blood pressure	10
4: Stationary or slight increase in blood pressure	15

- The lab then conducted a clinical trial with a random sample of 200 patients with high blood pressure.
- Use the sample data in the table below to test the hypothesis that the cell probabilities associated with the test preparation are identical to those for the standard. Use a significance level of 5%

Category	Observed Cell Counts
1	120
2	60
3	10
4	10

Contingency Tables: Test for Independence

Example

- The Centers for Disease Control and Prevention wants to determine if the severity of a skin disease is related to the age of the patient
- A patient's skin disease is classified as moderate, mildly severe, or severe
- The patients are divided into four age categories
- The distribution of skin disease over age categories is as follows

	Age Category				
Severity	I	II	III	IV	All Ages
Moderate	15	32	18	5	70
Mildly Severe	8	29	23	18	78
Severe	1	20	25	22	68
All Severities	24	81	66	45	216

- Do these data provide enough evidence for a significant relationship between the severity of skin disease and the patient's age?

Notation

- Assume we have two categorical variables: one factor has r levels, the other factor has c levels
- The frequency data are arranged in a cross tabulation, called **contingency table**, with r rows and c columns
- We denote the population proportion (or probability) falling in row i, column j as π_{ij}
- The total proportion for row i is $\pi_{i\bullet}$ and the total proportion for column j is $\pi_{\bullet j}$

Independence

- If the row and column proportions (probabilities) are independent, then $\pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}$
- The test statistic is the sum over all cells of (observed value - expected value)^2/expected value
- If $n_{i\bullet}$ (resp. $n_{\bullet j}$) is the actual frequency in row i (resp. column j), estimate $\pi_{i\bullet}$ (resp. $\pi_{\bullet j}$) by $\hat{\pi}_{i\bullet} = n_{i\bullet}/n$ (resp. $\hat{\pi}_{\bullet j} = n_{\bullet j}/n$)
- Under the hypothesis of independence, the estimated expected value in row i, column j is

$$\hat{E}_{ij} = n\hat{\pi}_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{n}$$

χ^2 Test of Independence

- H_0 : The two factors are independent vs. H_a : the two factors are dependent (associated)
- Test statistic: $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}}$
- For a specified significance level α, reject H_0 if χ^2 exceeds the critical value χ^2_α with $(r-1)(c-1)$ degrees of freedom
- In R, χ^2_α =`qchisq(1-alpha,(r-1)*(c-1))`

- In the example discussed at the beginning of the section, conduct a test to determine if the severity of the disease is independent of the age of the patient. Use a significance level of 5%

- Table of expected counts $\hat{E}_{ij}$

	Age Category			
Severity	I	II	III	IV
Moderate	7.78	26.25	21.39	14.58
Mildly Severe	8.67	29.2	23.83	16.25
Severe	7.56	25.50	20.78	14.17

χ^2 Test in R

```
skin.disease <- read.csv("skin_disease.csv", row.names = 1)
chisq.test(skin.disease)
```

```
    Pearson's Chi-squared test

data:  skin.disease
X-squared = 27.135, df = 6, p-value = 0.0001366
```

Validity of the Chi-square Test of Independence

- This test is based on an asymptotic approximation which requires a reasonably large sample size
- A conservative rule is that each $\hat{E}_{ij}$ must be at least 1 and no more than 20% of the $\hat{E}_{ij}$'s can be less than 5
- Standard practice when some of the $\hat{E}_{ij}$'s are too small is to combine those rows or columns with small totals until the rule is satisfied

Odds and Odds Ratios

Odds

- Another way to analyze count data on qualitative variables is to use the concept of odds (widely used in biomedical studies)
- The **odds** of an event A is $\dfrac{P(A)}{1-P(A)}$
- What are the odds of an event that has probability $2/3$ of happening?
- It is easy to convert the odds of an event back to the probability of the event
- If the odds of a horse (not winning) are stated as 9 to 1, what is the probability of the horse not winning?

Odds Ratios

- The **odds ratio** is the ratio of the odds of an event (for example, contracting a certain form of cancer) for one group (for example, men) to the odds of the same event for another group (for example, women)
- If A is any event with probabilities $P(A|\text{group 1})$ and $P(A|\text{group 2})$, the odds ratio (OR) is

$$OR = \frac{P(A|\text{group 1})/[1 - P(A|\text{group 1})]}{P(A|\text{group 2})/[1 - P(A|\text{group 2})]}$$

- The odds ratio equals 1 if the event A is statistically independent of group

Example

- A study was conducted to determine if the level of stress in a person's job affects his or her opinion about the company's proposed new health plan. A random sample of 3,000 employees yields the responses shown here

	Employee Response		
Job Stress	Favorable	Unfavorable	Total
Low	250	750	1,000
High	400	1,600	2,000
Total	650	2,350	3,000

- Estimate the conditional probabilities of a favorable and unfavorable response given the level of stress
- Compute an estimate of the odds ratio of a favorable response for the two groups