Lab 008 - Contingency Tables and Difference of two Proportions

EPIB607 - Inferential Statistics[a]

[a]Fall 2020, McGill University

This version was compiled on November 17, 2020

Contents

In the setting where a binary outcome is recorded for a single group of participants, inference about the binomial probability of success provides information about a population proportion p. Just as inference can be done for the difference of two population means, inference can also be done in the setting of comparing two population proportions p_1 and p_2. This lab then generalizes inference for binomial proportions to the setting of two-way contingency tables. Hypothesis testing in a two-way table assesses whether the two variables of interest are associated; this approach can be applied to settings with two or more groups and for responses that have two or more categories. Measures of association in two-by-two tables are also discussed.

1. Difference of two proportions

Recap: Inference for a single proportion

Suppose that X is binomial with parameters n (total number of trials) and p, the population parameter of success, where x represents the number of successes. Inference about p is based on the sample proportion $\hat{p}$, where $\hat{p} = x/n$; $\hat{p}$ is the point estimate of p.

Inference for p can be made using the normal approximation to the binomial, or directly using the binomial distribution.

- *Inference with the normal approximation*

  - The sampling distribution of $\hat{p}$ is approximately normal when 1) the sample observations are independent, 2) $np \geq 10$, $n(1-p) \geq 10$.[1] Under these conditions, the sampling distribution of $\hat{p}$ is approximately normally distributed with mean p and standard deviation

[1] The second condition is commonly referred to as the **success-failure condition**, since it can be effectively restated as the number of successes is greater than 10 and the number of failures is greater than 10.

$\sqrt{\frac{p(1-p)}{n}}$. For confidence intervals, substitute $\hat{p}$ for p; for hypothesis testing, substitute p_0 for p.

- The approximate two-sided 95% confidence interval for p is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- The test statistic z for the null hypothesis $H_0 : p = p_0$ based on a sample size of n is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{(p_0)(1-p_0)}{n}}}$$

- *Inference with exact methods*

  - Confidence intervals and p-values based on the binomial distribution are best calculated via R.

  - The logic behind calculating a p-value from the binomial distribution: Let X be a binomial random variable with parameters n and p_0, where $\hat{p} = x/n$ and x is the observed number of events. For a test of $H_0 : p = p_0$ versus $H_A : p \neq p_0$, the p-value equals $2 \times P(X \geq x)$.

Inference for the difference of two proportions

The normal model can be applied to $\hat{p}_1 - \hat{p}_2$ if the sampling distribution for each sample proportion is nearly normal, and if the samples are independent random samples from the relevant populations and independent of each other.

Each sample proportion approximately follows a normal model when $n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$, and $n_2(1-p_2)$ are all ≥ 10. To check success-failure in the context of a confidence interval, use $\hat{p}_1 and \hat{p}_2$.

The standard error of the difference in sample proportions is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

For hypothesis testing, an estimate of p is used to compute the standard error of $\hat{p}_1 - \hat{p}_2$: $\hat{p}$, the weighted average of the sample proportions $\hat{p}_1$ and $\hat{p}_2$,

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}.$$

To check success-failure in the context of hypothesis testing, check that $\hat{p}n_1$ and $\hat{p}n_2$ are both ≥ 10.

1. The use of screening mammograms for breast cancer has been controversial for decades because the overall benefit on breast cancer mortality is uncertain. A 30-year study to investigate the effectiveness of mammograms versus a standard non-mammogram breast cancer exam was conducted in Canada with 89,835 female participants.[2] Each woman was randomized to receive

	Death from breast cancer?	
	Yes	No
Mammogram Group	500	44,425
Control Group	505	44,405

either annual mammograms or standard physical exams for breast cancer over a 5-year screening period.

By the end of the 25 year follow-up period, 1,005 women died from breast cancer. The results are summarized in the following table.[3]

a) Calculate $\hat{p}_1$ and $\hat{p}_2$, the two sample proportions of interest.

b) Analyze the results; do the data suggest that annual mammography results in a reduction in breast cancer mortality relative to standard exams? Be sure to check the assumptions for using the normal approximation.

c) Calculate and interpret a 95% confidence interval for the difference in proportions of deaths from breast cancer. Be sure to check the assumptions for using the normal approximation.

2. Remdesivir is an antiviral drug previously tested in animal models infected with coronaviruses like SARS and MERS. As of May 2020, remdesivir had temporary approval from the FDA for use in severely ill COVID-19 patients and was the subject of numerous ongoing studies.

   A randomized controlled trial conducted in China enrolled 236 patients with severe COVID-19; 158 were assigned to receive remdesivir and 78 to receive a placebo. In the remdesivir group, 103 patients showed clinical improvement; in the placebo group, 45 patients showed clinical improvement.[4]

   a) Calculate $\hat{p}_1$ and $\hat{p}_2$, the two sample proportions of interest.

   b) Conduct a formal comparison of the clinical improvement rates and summarize your findings. Be sure to check the assumptions for using the normal approximation.

   c) Report and interpret an appropriate interval estimate. Be sure to check the assumptions for using the normal approximation.

[2] Miller AB. 2014. *Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial. BMJ* 348 (2014): g366.

[3] During the 25 years following the screening period, each woman was screened for breast cancer according to the standard of care at her health care center.

[4] Wang, Y, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multi-centre trial. *Lancet* 395(10236). 16 May 2020.

2. Contingency Tables

The χ^2 test of independence

In the χ^2 test of independence, the observed number of cell counts are compared to the number of **expected** cell counts, where the expected counts are calculated under the null hypothesis.

- H_0: the row and column variables are not associated
- H_A: the row and column variables are associated

The expected count for the i^{th} row and j^{th} column is

$$E_{i,j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{n},$$

where n is the total number of observations.

Assumptions for the χ^2 test:

- *Independence*. Each case that contributes a count to the table must be independent of all other cases in the table.
- *Sample size*. Each expected cell count must be greater than or equal to 10. For tables larger than 2×2, it is appropriate to use the test if no more than 1/5 of the expected counts are less than 5, and all expected counts are greater than 1.

The χ^2 **test statistic** is calculated as

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}},$$

and is approximately distributed χ^2 with degrees of freedom $(r-1)(c-1)$, where r is the number of rows and c is the number of columns. $O_{i,j}$ represents the observed count in row i, column j.

For each cell in a table, the **residual** equals

$$\frac{O_{i,j} - E_{i,j}}{\sqrt{E_{i,j}}}.$$

Residuals with a large magnitude contribute the most to the χ^2 statistic. If a residual is positive, the observed value is greater than the expected value; if a residual is negative, the observed value is less than the expected.

Fisher's exact test

When the expected counts in a two-way table are less than 10, Fisher's exact test is used to compute a p-value without relying on the normal approximation. In this course, only the logic behind Fisher's exact test for a 2×2 table is discussed. In the 2×2 table case, the hypotheses for Fisher's exact test can be expressed in the same way as for a two-sample test of proportions; the null hypothesis is $H_0 : p_1 = p_2$.

The p-value is the probability of observing results as or more extreme than those observed under the assumption that the null hypothesis is true.

- Thus, the p-value is calculated by adding together the individual conditional probabilities of obtaining each table that is as or more extreme than the one observed, under the null hypothesis and given that the marginal totals are considered fixed.
- When the marginal totals are held constant, the value of any one cell in the table determines the rest of entries. When marginal totals are considered fixed, each table represents a unique set of results.
- Extreme tables are those which contradict $H_0 : p_1 = p_2$.
- A two-sided p-value can be calculated by doubling the smaller of the possible one-sided p-values; this method is typically used when calculating p-values by hand. Another common method is to classify "more extreme" tables as all tables with probabilities less than that of the observed table, in both directions; the p-value is the sum of probabilities for the qualifying table.

The probability of a particular table (i.e., set of results) can be calculated with the **hypergeometric distribution**.

Let X represent the number of successes in a series of repeated Bernoulli trials, where sampling is done without replacement. Suppose that in the population of size N, there are m total successes. What is the probability of observing exactly k successes when drawing a sample of size n?

For example, imagine an urn with m white balls and $N-m$ black balls (thus, there are N total balls). Draw n balls without replacement (i.e., a sample of n balls). What is the probability of observing k white balls in the sample?

The possible results of a sample can be organized in a 2×2 table:

	White Ball	Black Ball	Total
Sampled	k	$n-k$	n
Not Sampled	$m-k$	$N-n-(m-k)$	$N-n$
Total	m	$N-m$	N

The probability of observing exactly k successses in a sample of size n (i.e., n dependent trials) is given by

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}.$$

Hypergeometric probabilities are calculated in R with the use of `dhyper()` and `phyper()`. The following code shows how to calculate $P(X = 5)$, $P(X \leq 5)$, and $P(X > 5)$ for $X \sim \text{HGeom}(10, 15, 8)$, where $m = 10$, $N - m = 15$, and $n = 8$.

```
#probability X equals 5
dhyper(5, 10, 15, 8)
```

```
#  [1] 0.1060121
```

```
#probability X is less than or equal to 5
phyper(5, 10, 15, 8)
```

```
#  [1] 0.9779072
```

```
#probability X is greater than 5
phyper(5, 10, 15, 8, lower.tail = FALSE)
```

```
#  [1] 0.02209278
```

Measures of association in two-by-two tables

Chapter 1 introduced the **relative risk (RR)**, a measure of the risk of a certain event occurring in one group relative to the risk of the event occurring in another group, as a numerical summary for two-by-two (2×2) tables. The relative risk can also be thought of as a measure of association.

Consider the following hypothetical two-by-two table. The relative risk of Outcome A can be calculated by using either Group 1 or Group 2 as the reference group:

	Outcome A	Outcome B	Sum
Group 1	a	b	$a+b$
Group 2	c	d	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d=n$

Table 1. A hypothetical two-by-two table of outcome by group.

$$RR_{\text{A, comparing Group 1 to Group 2}} = \frac{a/(a+b)}{c/(c+d)}$$

$$RR_{\text{A, comparing Group 2 to Group 1}} = \frac{c/(c+d)}{a/(a+b)}$$

The relative risk is only valid for tables where the proportions $a/(a+b)$ and $c/(c+d)$ represent the incidence of Outcome A within the populations from which Groups 1 and 2 are sampled.

The **odds ratio (OR)** is a measure of association that remains applicable even when it is not possible to estimate incidence of an outcome from the sample data. The **odds** of Outcome A in Group 1 are a/b, while the odds of Outcome A in Group 2 are c/d.

$$OR_{\text{A, comparing Group 1 to Group 2}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$OR_{\text{A, comparing Group 2 to Group 1}} = \frac{c/d}{a/b} = \frac{bc}{ad}$$

3. The χ^2 test of independence

3. In resource-limited settings, single-dose nevirapine (NVP) is given to an HIV-positive woman during birth to prevent mother-to-child transmission of the virus. Exposure of the infant to NVP may foster the growth of more virulent strains of the virus in the child.

   If a child is HIV-positive, should they be treated with NVP or a more expensive drug, lopinavir (LPV)? In this setting, success means preventing a growth of the virus in the child (i.e., preventing

virologic failure). The following table contains data from a 2012 study conducted in six African countries and India.[5]

	NVP	LPV	Total
Virologic Failure	60	27	87
Stable Disease	87	113	200
Total	147	140	287

a) State the null and alternative hypotheses.

b) Calculate the expected cell counts.

c) Check the assumptions for using the χ^2 test.

d) Calculate the χ^2 test statistic.

e) Calculate the p-value for the test statistic using `pchisq()`. The p-value represents the probability of observing a result as or more extreme than the sample data.

f) Confirm the results from parts c) and d) using `chisq.test()`. Note that the value of the test statistic will be slightly different because R is applying a 'continuity correction'.

g) Summarize the conclusions; be sure to include which drug is recommended for treatment, based on the data.

h) Repeat the analysis using inference for the difference of two proportions and confirm that the results are the same.

4. In the PREVEND study introduced in Chapter 6, researchers measured various features of study participants, including data on statin use and highest level of education attained. From the data in `prevend.samp`, is there evidence of an association between statin use and educational level? Summarize the results.

4. Fisher's exact test

5. *Clostridium difficile* is a bacterium that causes inflammation of the colon. Antibiotic treatment is typically not effective, particularly for patients who experience multiple recurrences of infection. Infusion of feces from healthy donors has been reported as an effective treatment for recurrent infection. A randomized trial was conducted to compare the efficacy of donor-feces infusion versus vancomycin, the antibiotic typically prescribed to treat *C. difficile* infection. The results of the trial are shown in the following table.

	Cured	Uncured	Sum
Fecal Infusion	13	3	16
Vancomycin	4	9	13
Sum	17	12	29

[5] A. Violari, et al. "Nevirapine versus ritonavir-boosted lopinavir for HIV-infected children." *NEJM* 366: 2380-2389.

a) Can a χ^2 test be used to analyze these results?

b) Researchers are interested in understanding whether fecal infusion is a more effective treatment than vancomycin. Write the null hypothesis and appropriate one-sided alternative hypothesis.

c) Under the assumption that the marginal totals are fixed, enumerate all possible sets of results that are more extreme than what was observed, in the same direction.

d) Calculate the probability of the observed results.

e) Calculate the probability of each set of results enumerated in part c).

f) Based on the answers in parts d) and e), compute the one-sided p-value and interpret the results.

g) Use `fisher.test( )` to confirm the calculations in part f) and to calculate the two-sided p-value.

6. Psychologists conducted an experiment to investigate the effect of anxiety on a person's desire to be alone or in the company of others (Schacter 1959; Lehmann 1975). A group of 30 individuals were randomly assigned into two groups; one group was designated the "high anxiety" group and the other the "low anxiety" group. Those in the high-anxiety group were told that in the "upcoming experiment", they would be subjected to painful electric shocks, while those in the low-anxiety group were told that the shocks would be mild and painless.[6]. All individuals were informed that there would be a 10 minute wait before the experiment began, and that they could choose whether to wait alone or with other participants.

   The following table summarizes the results:

	Wait Together	Wait Alone	Sum
High-Anxiety	12	5	17
Low-Anxiety	4	9	13
Sum	16	14	30

a) Under the null hypothesis of no association, what are the expected cell counts?

b) Under the assumption that the marginal totals are fixed and the null hypothesis is true, what is the probability of the observed set of results?

c) Enumerate the tables that are more extreme than what was observed, in the same direction.

d) Conduct a formal test of association for the results and summarize your findings. Let $\alpha = 0.05$.

[6] Individuals were not actually subjected to electric shocks of any kind

5. Measures of association in two-by-two tables

7. Suppose a study is conducted to assess the association between smoking and cardiovascular disease (CVD). Researchers recruited a group of 231 study participants then categorized them according to smoking and disease status: 111 are smokers, while 40 smokers and 32 non-smokers have CVD. Calculate and interpret the relative risk of CVD.

8. Suppose another study is conducted to assess the association between smoking and CVD, but researchers use a different design: 90 individuals with CVD and 110 individuals without CVD are recruited. 40 of the individuals with CVD are smokers, and 80 of the individuals without CVD are non-smokers.

    a) Is relative risk an appropriate measure of association for these data? Explain your answer.

    b) Calculate the odds of CVD among smokers and the odds of CVD among non-smokers.

    c) Calculate and interpret the odds ratio of CVD, comparing smokers to non-smokers.

    d) What would an odds ratio of CVD (comparing smokers to non-smokers) equal to 1 represent, in terms of the association between smoking and CVD? What would an odds ratio of CVD less than 1 represent?