

Unit 8: Inference for Categorical Data

Statistics 102 Teaching Team

April 20, 2020

Introduction

Inference for binomial proportions

Inference for two-way tables

Measures of association in two-by-two tables

Introduction

TOOLS FOR ASSESSING ASSOCIATION

So far, we have covered methods for numerical outcomes.

- numerical outcome with a categorical predictor
 - two-sample t -tests and ANOVA
- numerical outcome with a numerical (or categorical) predictor
 - simple linear regression
- numerical outcome with several predictors, numerical or categorical
 - multiple linear regression

TOOLS FOR ASSESSING ASSOCIATION...

Next, we will cover methods for categorical outcomes.

- categorical outcome with a categorical predictor
 - χ^2 test of independence in a two-way table
- binary outcome with a numerical (or categorical) predictor
 - simple logistic regression
- binary outcome with several predictors, numerical or categorical
 - multiple logistic regression

Inference for binomial proportions

ADVANCED MELANOMA

Advanced melanoma is an aggressive form of skin cancer that until recently was almost uniformly fatal.

Research is being conducted on therapies that might trigger an immune response to the cancer and cause the melanoma to stop progressing or disappear entirely.

In a study where 52 patients were treated concurrently with two new therapies, nivolumab and ipilimumab, 21 had an immune response.¹

¹Wolchok, et. al. *NEJM* (2013) 369(2): 122-33.

ADVANCED MELANOMA. . .

Questions that can be addressed with inference. . .

- What is the estimated population probability of immune response following concurrent therapy with nivolumab and ipilimumab?
- What is the 95% confidence interval for the estimated population probability of immune response following concurrent therapy with nivolumab and ipilimumab?
- In previous studies, the proportion of patients responding to one of these agents was 30% or less. Do these results suggest that the probability of response to concurrent therapy is better than 0.30?

INFERENCE FOR BINOMIAL PROPORTIONS

The melanoma data are binomial data, with success defined as experiencing an immune response.

Suppose X is a binomial random variable with parameters n and p , where n is the number of trials and p is the probability of success.

- Inference is made about the population parameter p , the probability of success in the population.
- The estimate of p from the observed sample is $\hat{p} = x/n$, where x is the observed number of successes.

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

ASSUMPTIONS FOR USING THE NORMAL DISTRIBUTION

The sampling distribution of \hat{p} is approximately normal when

1. The sample observations are independent, and
2. At least 10 successes and 10 failures are expected in the sample: $np \geq 10$ and $n(1 - p) \geq 10$.²

Under these conditions, \hat{p} is approximately normally distributed with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

Since p is unknown, it is necessary to substitute either \hat{p} or p_0 for p when using the standard error of \hat{p} to compute confidence intervals and conduct hypothesis tests.

²This condition is commonly referred to as the success-failure condition.

INFERENCE WITH THE NORMAL APPROXIMATION

In the context of calculating CIs, substitute \hat{p} for p .

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

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

```
#calculate confidence interval  
prop.test(x = 21, n = 52, conf.level = 0.95)$conf.int
```

```
## [1] 0.2731269 0.5487141  
## attr(,"conf.level")  
## [1] 0.95
```

INFERENCE WITH THE NORMAL APPROXIMATION...

In the context of hypothesis testing, substitute p_0 for p .

The test statistic z for the null hypothesis $H_0 : p = p_0$ is

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

```
#conduct hypothesis test  
prop.test(x = 21, n = 52, p = 0.30, alternative = "greater")
```

```
##  
## 1-sample proportions test with continuity correction  
##  
## data: 21 out of 52, null probability 0.3  
## X-squared = 2.1987, df = 1, p-value = 0.06906  
## alternative hypothesis: true p is greater than 0.3  
## 95 percent confidence interval:  
## 0.2906582 1.0000000  
## sample estimates:  
## p  
## 0.4038462
```

EXACT INFERENCE FOR BINOMIAL DATA

```
#use pbinom  
pbinom(20, 52, p = 0.30, lower.tail = FALSE)
```

```
## [1] 0.07167176
```

```
#use binom.test  
binom.test(x = 21, n = 52, p = 0.30, alternative = "greater")
```

```
##  
## Exact binomial test  
##  
## data: 21 and 52  
## number of successes = 21, number of trials = 52, p-value = 0.07167  
## alternative hypothesis: true probability of success is greater than 0.3  
## 95 percent confidence interval:  
## 0.2889045 1.0000000  
## sample estimates:  
## probability of success  
## 0.4038462
```

INFERENCE FOR THE DIFFERENCE OF TWO PROPORTIONS

The normal model can be applied to $\hat{p}_1 - \hat{p}_2$ if

1. The two samples are independent, the observations in each sample are independent, and
2. At least 10 successes and 10 failures are expected in each sample.

The standard error of the difference in sample proportions is

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

In hypothesis testing, the following estimate of p is used to compute the standard error:

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

EFFECTIVENESS OF MAMMOGRAMS

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. Each woman was randomized to receive 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.

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

EFFECTIVENESS OF MAMMOGRAMS...

```
#analyze the data
```

```
prop.test(x = c(500, 505), n = c(500 + 44425, 505 + 44405))
```

```
##  
## 2-sample test for equality of proportions with continuity correction  
##  
## data:  c(500, 505) out of c(500 + 44425, 505 + 44405)  
## X-squared = 0.01748, df = 1, p-value = 0.8948  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.001512853 0.001282751  
## sample estimates:  
##      prop 1      prop 2  
## 0.01112966 0.01124471
```


Inference for two-way tables

INFERENCE FOR TWO-WAY TABLES

A two-way table summarizes information about the relationship between two categorical variables.

Testing for a difference between p_1 and p_2 is equivalent to testing for association in a two-way table that has two rows and two columns.

TREATING HIV⁺ INFANTS

In resource-limited settings, single-dose nevirapine is given to an HIV⁺ woman during birth to prevent mother-to-child transmission of the virus.

- Exposure of the infant to nevirapine (NVP) may foster the growth of resistant strains of the virus in the child.
- If the child is HIV⁺, should he/she be treated with nevirapine or a more expensive drug, lopinavir (LPV)?

In this setting, the possible outcomes are virologic failure (the virus becomes resistant) versus stable disease (virus growth is prevented).

TREATING HIV⁺ INFANTS...

The following table summarizes the results of a 2012 study comparing NVP versus LPV in treatment of HIV-infected infants.³ Children were randomized to receive either NVP or LPV.

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

³Violari, et al. *NEJM* 366: 2380-2389.

FORMULATING HYPOTHESES IN A TWO-WAY TABLE

The main question of interest:

- Do the data support the claim of a difference in outcome by treatment?

If there is no difference in outcome by treatment, then knowing treatment provides no information about outcome; treatment assignment and outcome are *independent* (i.e., *not associated*).

- H_0 : Treatment and outcome are not associated.
- H_A : Treatment and outcome are associated.
 - This is inherently a two-sided alternative.

THE χ^2 TEST OF INDEPENDENCE

In the (Pearson) χ^2 test, 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.

The test statistic quantifies how far the observed results deviate from what is expected under the null hypothesis.

- A large test statistic represents stronger evidence against the null hypothesis of independence.

EXPECTED CELL COUNTS

If treatment had no effect on outcome, what would we expect to see?

- Let $A = \{\text{assignment to NVP}\}$
- Let $B = \{\text{virologic failure}\}$

Under the hypothesis of independence,

$$P(A \text{ and } B) = P(A) \times P(B) = \left(\frac{147}{287}\right) \left(\frac{87}{287}\right)$$

The expected cell count in the upper left corner would be

$$(287) \left(\frac{147}{287}\right) \left(\frac{87}{287}\right) = 44.56$$

What about the other cells?

FORMULA FOR EXPECTED CELL COUNTS

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

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 .

THE χ^2 TEST IN R

```
hiv.table = matrix(c(60, 27, 87, 113), nrow = 2, ncol = 2, byrow = T)
dimnames(hiv.table) = list("Outcome" = c("V. Failure", "Stable Disease"),
                           "Drug" = c("NVP", "LPV"))
chisq.test(hiv.table)
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  hiv.table
## X-squared = 14.733, df = 1, p-value = 0.0001238
```

```
chisq.test(hiv.table)$expected
```

```
##
##      Drug
## Outcome      NVP      LPV
## V. Failure    44.56098 42.43902
## Stable Disease 102.43902 97.56098
```

RESIDUALS IN THE χ^2 TEST

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.

RESIDUALS IN `prevend.samp`

```
chisq.test(prevend.samp$Statin, prevend.samp$Education)
```

```
##
##  Pearson's Chi-squared test
##
## data:  prevend.samp$Statin and prevend.samp$Education
## X-squared = 19.054, df = 3, p-value = 0.0002665
```

```
chisq.test(prevend.samp$Statin, prevend.samp$Education)$residuals
```

```
##
##               prevend.samp$Education
## prevend.samp$Statin  Primary  LowerSec  UpperSec    Univ
##               NonUser -1.3196995 -0.8994999  0.3760673  1.3000955
##               User    2.4146629  1.6458208 -0.6880929 -2.3787932
```

Measures of association in two-by-two tables

RELATIVE RISK IN A 2×2 TABLE

The **relative risk (RR)** is a measure of the risk of a certain event occurring in one group relative to the risk of the event occurring in another group.

The risk of virologic failure among the NVP group is

$$\frac{\# \text{ in NVP group and had virologic failure}}{\text{total } \# \text{ in NVP group}} = \frac{60}{147} = 0.408$$

The risk of virologic failure among the LPV group is

$$\frac{\# \text{ in LPV group and had virologic failure}}{\text{total } \# \text{ in LPV group}} = \frac{27}{140} = 0.193$$

Thus, the relative risk of virologic failure comparing NVP to LPV is $0.408/0.193 = 2.11$.

- Children treated with NVP are estimated to be more than twice as likely to experience virologic failure.

THE ODDS RATIO IN A 2×2 TABLE

The **odds ratio (OR)** is a measure of the odds of a certain event occurring in one group relative to the risk of the event occurring in another group.

The odds of virologic failure among the NVP group is

$$\frac{\# \text{ in NVP group and had virologic failure}}{\# \text{ in NVP group and did not have virologic failure}} = \frac{60}{87} = 0.690$$

The odds of virologic failure among the LPV group is

$$\frac{\# \text{ in LPV group and had virologic failure}}{\# \text{ in LPV group and did not have virologic failure}} = \frac{27}{113} = 0.239$$

Thus, the odds ratio of virologic failure comparing NVP to LPV is $0.690/0.239 = 2.89$.

- The odds of virologic failure when treated with NVP are almost three times as large as the odds of virologic failure when treated with LPV.

RELATIVE RISK VERSUS ODDS RATIO

The relative risk cannot be used in studies that use **outcome-dependent sampling**, such as a case-control study:

- Suppose in the HIV study, researchers had identified 100 HIV-positive infants who had experienced virologic failure (cases) and 100 who had stable disease (controls), then recorded the number in each group who had been treated with NVP or LPV.
- With this design, the sample proportion of infants with virologic failure no longer estimates the population proportion.
 - Similarly, the sample proportion of infants with virologic failure in a treatment group no longer estimates the proportion of infants who would experience virologic failure in a hypothetical population treated with that drug.

The odds ratio remains valid even when it is not possible to estimate incidence of an outcome from sample data.