Class Notes: STAT 501
Nonparametrics & Log-Linear Models
Comparing Proportions in 2x2 Tables
Chi-squared Test

Da Kuang
University of Pennsylvania

Contents

1 Proportions in Table

In this chapter, the object is to compare two unknown success probabilities,p_1, p_2, on the basis of the corresponding rates of success in independent samples. We assume the sample size is large enough so that we could apply the central limit theorem.

1.1 Background

We observe the outcomes of n_1 independent repeated Bernoulli trials, each with success probability p_1. We also observe the outcomes of n_2 independent repeated Bernoulli trials, each with success probability p_2.

	Successes	Failures	Totals
Sample 1	$\mathcal{O}_{11}$	$\mathcal{O}_{12}$	$n_{1\cdot}$
Sample 2	$\mathcal{O}_{21}$	$\mathcal{O}_{22}$	$n_{2\cdot}$
Totals:	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

Figure 1: 2×2 Table of Outcomes

1.2 Conditional Test vs. Unconditional Test

In many 2×2 tables, the row of a table refer to different groups, the sample size for those groups are often fixed by the sample design. Sometimes we refer rows as samples, or X and refer columns as results, or Y.

When the marginal totals for the levels of X (Samples) are fixed rather than random, a joint distribution for X and Y (Result) is not meaningful. We are actually more interested in the conditional distribution, $\mathbf{P}(Y = 1|X = 1)$.

As there are two outcome categories for Y , the binomial distribution applies for each

conditional distribution. We assume a binomial distribution for the sample in each row, with number of trials equal to the fixed row total.

When there are more than two outcome categories for Y , such as (*always, sometimes, never*), the multinomial distribution applies for each conditional distribution.

To better understand this, let us consider a generic 2×2 contingency table obtained by inoculating 15 subjects with vaccine and 15 subjects with placebo.

Treatment	Infection Status		Total
	Yes	No	
Vaccine	x_e	$15 - x_e$	15
Placebo	x_c	$15 - x_c$	15
Total	$x_e + x_c$	$30 - x_e - x_c$	30

From unconditional test's perspective, the probability of observing the table is a product of two binomials. Suppose the null hypothesis is $\pi_e = \pi_c = \pi$.

$$\mathbf{P}(\text{Observe the Table}|\pi) = \binom{15}{x_c}\binom{15}{x_e}\pi^{x_c+x_e}(1-\pi)^{30-x_c-x_e}$$

The p-value is

$$p = \sum_{T(\text{Observed Table}) \geq T(\text{Null Table})} \mathbf{P}(\text{Observe the Table}|\pi)$$

From the conditional test's perspective,the probability of observing the table is a product of two binomials. Suppose we observe 19 infections in total.

$$\mathbf{P}(\text{Observe the Table}|x_c + x_e = 19) = \frac{\binom{15}{x_c}\binom{15}{x_e}}{\binom{30}{19}}$$

The p-value is

$$p = \sum_{T(\text{Observed Table}) \geq T(\text{Null Table})} \mathbf{P}(\text{Observe the Table} | x_c + x_e = 19)$$

1.3 Assumption

It has been shown that conditional test outperform unconditional test. So here we make n_1 and n_2 fixed then have the following assumptions.

- O_{11} is the number of success observed in $n_{1\cdot}$ independent Bernoulli trials, each with success probability π_1.
- O_{21} is the number of success observed in $n_{2\cdot}$ independent Bernoulli trials, each with success probability π_2.
- The Bernoulli trials corresponding to sample 1 are independent of the Bernoulli trials corresponding to sample 2.

1.4 Hypothesis Test

Given the following 2×2 contingency, it is known that n_1 and n_2 are fixed. Then n_{11}, $n_1 - n_{11}$, n_{21}, and $n_2 - n_{21}$ are random variables.

	Success	Failure	Total
Treatment	n_{11}		n_1
Control	n_{21}		n_2

Figure 2: Random Variable in Contingency Table

The sample proportions are $p_1 = n_{11}/n_1$ and $p_2 = n_{21}/n_2$.

Since $n_{11} \sim \text{Binomial}(n_1, \pi_1)$, we have

$$\mathbf{E}p_1 = \pi_1, \mathbf{Var}p_1 = \frac{\pi_1(1-\pi_1)}{n_1}$$

Since $n_{21} \sim \text{Binomial}(n_2, p_2)$, we have

$$\mathbf{E}p_2 = \pi_2, \mathbf{Var}p_2 = \frac{\pi_2(1-\pi_2)}{n_2}$$

Therefore,

$$\mathbf{E}(p_2 - p_1) = \mathbf{E}p_2 - \mathbf{E}p_1 = \pi_2 - \pi_1$$

Interestingly, given n_1 and n_2, the derived sample porpotion can be used to estimate the unobservable probability. The null hypothesis is constructed based on that.

1.4.1 Approximation

The hypothesis of interest is

$$H_0 : \pi_1 = \pi_2 = \pi$$

with the common value π being unspecified.

We construct the test statistic based on the difference between p_1 and p_2. For Wald test, we need the mean and standard deviation of $p_1 - p_2$.

- Based on null hypothesis, mean$(p_1 - p_2) = 0$.
- For SD, suppose $p = \frac{n_{11}+n_{21}}{n_1+n_2}$,

$$\text{SD}(p_1 - p_2) = \sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}$$

Therefore, we have the Wald test statistic of $p_1 - p_2$,

$$T = \frac{p_1 - p_2}{\text{SD}(p_1 - p_2)}$$

The alternative hypothesis and reject region are

- $H_1 : \pi_1 > \pi_2$: Reject H_0 if $T \geq z_\alpha$.
- $H_1 : \pi_1 < \pi_2$: Reject H_0 if $T \leq z_\alpha$.
- $H_1 : \pi_1 \neq \pi_2$: Reject H_0 if $|T| \geq z_{\alpha/2}$

The construction of the confidence interval for $\pi_1 - \pi_2$ is a little bit different than usual one-sample binomial distribution, where CI can be calculated by inverting the test statistic because of the duality.

The test statistic T is based on the null hypothesis, where the two binomial distribution share the same p, while we see them as two different distribution when estimating the CI, so that we have

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

Hence, the confidence interval for $\pi_1 - \pi_2$ is

$$p_1 - p_2 - Z_{\alpha/2}S \leq \pi_1 - \pi_2 \leq p_1 - p_2 + Z_{\alpha/2}S$$

1.4.2 Notes on Alternative Hypothesis

When choosing alternative hypothesis, by convention people choose two-sided-test, it is harder to be rejected. In other words, if the two-sided alternative hypothesis is rejected then the one-sided must be rejected.

Usually, people first use two-sided hypothesis to reject the null hypothesis, then choose the side by infering the relationship based on the data.

1.4.3 Notes on Continuity Correction

Whether or not to apply continuity correction is debatable. Be convention, if you decide to use continuity correction, you must report it in the paper.

If the data size is large, the continuity correction usually have no effect on the rejection of hypothesis. But when the data is small, you may reach the different conclusion of rejection. Then we need to do further simulation to exam the model.

1.5 Action in R

Recall that the square of normal distribution is the chi-squared distribution with one degree of freedom. So the following two calculations give the same p-values.

```
n1=60; n2=40; n=n1+n2; n11=20; n21=30;
p1=n11/n1; # 0.3333333
p2=n21/n2; # 0.75
p=(n11+n21)/n; # 0.5
stat=(p1-p2)/sqrt(p*(1-p)/n1+p*(1-p)/n2)
stat # -4.082483
2*pnorm(stat) # 4.455709e-05
stat^2 # 16.66667
pchisq(stat^2,1,lower.tail=F) # 4.455709e-05
```

Probabilities of success in R without continuity correction.

```
s=sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2) # 0.09160351
z=qnorm(0.975) # 1.959964
# Confidence Internval
p1-p2+c(-1,1)*z*s
# [1] -0.5962063 -0.2371271

```

```r
prop.test(c(n11,n21),c(n1,n2),correct=F)
# 2-sample test for equality of proportions without continuity correction
# data: c(n11, n21) out of c(n1, n2)
# X-squared = 16.667, df = 1, p-value = 4.456e-05
# alternative hypothesis: two.sided
# 95 percent confidence interval:
# -0.5962063 -0.2371271
# sample estimates:
# prop 1 prop 2
# 0.3333333 0.7500000
```

Probabilities of success in R with continuity correction.

```r
yates=min(0.5,abs(p1-p2)/(1/n1+1/n2))
# Confidence Internval
p1-p2+c(-1,1)*(z*s+yates*(1/n1+1/n2))
# [1] -0.6170396 -0.2162937
prop.test(c(n11,n21),c(n1,n2))
# 2-sample test for equality of proportions with continuity correction
# data: c(n11, n21) out of c(n1, n2)
# X-squared = 15.042, df = 1, p-value = 0.0001052
# alternative hypothesis: two.sided
# 95 percent confidence interval:
# -0.6170396 -0.2162937
# sample estimates:
# prop 1 prop 2
# 0.3333333 0.7500000
```

2 Chi Squared Test

In this section, we will talk about Chi Squared Test for homogeneity and for independence. The two tests are the same but the setups are different. In reality, statistics should get involved into the research as early as possible. Nowdays, people sometimes looks for the data and then try to apply statistic models on it because the computational cost is low. However, in the early age of the study, statistics can get involved with the experiment design.

2.1 Pearson's χ^2 Test for Multinomial Data

Pearson's χ^2 test is used for multinomial data. Recall that if $X = (X_1, \cdots, X_k)$ has a multinomial(n, p) distribution, then the MLE of p is $\widehat{p} = (\widehat{p_1}, \cdots, \widehat{p_k}) = (X_1/n, \cdots, X_k/n)$.

Let $p_0 = (p_{01}, \cdots, p_{0k})$ be some fixed vector and suppose we want to test

$$H_0 : p = p_0 \text{ versus } H_1 : p \neq p_0$$

Definition 2.1 (Pearson's χ^2 statistic) *Pearson's χ^2 statistic is*

$$T = \sum_{j=1}^{k} \frac{(X_j - np_{0j})^2}{np_{0j}} = \sum_{j=1}^{k} \frac{(X_j - E_j)^2}{E_j}$$

where $E_j = \mathbf{E}(X_j) = np_{0j}$ is the expected value of X_j under H_0.

Theorem 2.1 *Under H_0, $T \to \chi^2_{k-1}$. Hence the test: reject H_0 if $T > \chi^2_{k-1,\alpha}$ has asymptotic level α. The p-value is $\mathbf{P}(\chi^2_{k-1} > t)$ where t is the observed value of the test statistic.*

2.2 Test of Homogeneity

2.2.1 Assumptions

We have two populations and the sample sizes are fixed. The outcome is random within each population.

2.2.2 Hypothesis Test

The homogeneity hypothesis: the chance of success is the same for both populations.

The large-sample two-sided test based on T can also be presented via Karl Pearson's chi-squared statistic. T is used to check the difference between p_1 and p_2, to infer the relationship between π_1 and π_2. Now we are going to test the difference between the expected table and the observed table.

Suppose the null hypothesis, $H_0 : \pi_1 = \pi_2$, is true, then the best estimator of the common success probability is $p = n_{\cdot 1}/n$. Using this estimator, the expected values of the random quantities O_{11}, O_{12}, O_{21}, and O_{22} can be estimated, respectively by E_{11}, E_{12}, E_{21}, and E_{22}, where

$$
\begin{aligned}
E_{11} &= n_{1\cdot} \times p = \frac{n_1 n_{\cdot 1}}{n}; \\
E_{12} &= n_1 \times (1-p) = \frac{n_1 n_{\cdot 2}}{n}; \\
E_{21} &= n_2 \times p = \frac{n_2 n_{\cdot 1}}{n}; \\
E_{22} &= n_2 \times (1-p) = \frac{n_2 n_{\cdot 2}}{n}.
\end{aligned}
$$

For example, if $H_0 : \pi_1 = \pi_2 = 0.5$ is true, then the expected numbers are

	Death	
	Yes	No
Treatment	30	30
Control	20	20

Figure 3: The Exptected Contigency Table

A measure of the discrepancy between the observed frequencies and the estimated exptected frequencies under the hypothesis is the chi-squared statistic

$$S = \frac{(n_{11} - E_{11})^2}{E_{11}} + \frac{(n_{12} - E_{12})^2}{E_{12}} + \frac{(n_{21} - E_{21})^2}{E_{21}} + \frac{(n_{22} - E_{22})^2}{E_{22}}$$

Intuitively, the differences "observed $-$ expected", are squared, eliminating the balancing out of positive and negative discrepancies. Each squared difference is weighted by the inverse of the corresponding E, so that the difference involving small E's assume the greatest importance.

There is a shortcut formula for the calculation of the chi-squared test statistic, namely,

$$S = \frac{n(n_{11}n_{22} - n_{21}n_{12})^2}{n_{.1}n_{.2}n_1n_2}$$

It can be shown that $T^2 = S$. $S \sim \chi^2_1$, chi-squared distribution with 1 degree of freedom. This implies that the two-sided approximate α-level test of $p_1 - p_2 = 0$ versus $p_1 - p_2 \neq 0$, given by the reject region of $|T| \geq z_{\alpha/2}$, is equivalent to the test

- H_0: $\pi_1 = \pi_2$.
- H_1: $\pi_1 \neq \pi_2$. Reject H_0 if $S > \chi^2_{\alpha,1}$.

2.2.3 Action in R

`chisq.test` is called within `prop.test`.

```
n.1=n11+n21; p=n.1/n
nv=c(n11,n1-n11,n21,n2-n21)
ev=c(n1*p,n1*(1-p),n2*p,n2*(1-p))
cstat=sum((abs(nv -ev) - yates)^2/ev)
pchisq(cstat, 1, lower.tail = F)
# [1] 0.0001051636
prop.test(c(n11,n21),c(n1,n2))
# 2-sample test for equality of proportions with continuity correction
#
# data: c(n11, n21) out of c(n1, n2)
# X-squared = 15.042, df = 1, p-value = 0.0001052
# alternative hypothesis: two.sided
# 95 percent confidence interval:
# -0.6170396 -0.2162937
# sample estimates:
# prop 1 prop 2
# 0.3333333 0.7500000
xm=matrix(c(n11,n1-n11,n21,n2-n21),2,2)
chisq.test(xm)
# Pearson's Chi-squared test with Yates' continuity correction
#
# data: xm
# X-squared = 15.042, df = 1, p-value = 0.0001052
```

2.3 Test of Independence

2.3.1 Assumptions

The main difference in the set-up is that we do not random assign a subject to any group before collecting the data. So that we only have one population, instead of two, but we have four combinations of the result. Therefore, it is meaningful to consider the unconditional probability, i.e. the joint distribution of two factors.

- Total sample size n is fixed.
- Each observation from a general population is cross-classified on the basis of two factors X and Y.
- Cell counts are random.

	Belief in Afterlife	
Belief in Aliens	Yes	No
Yes	50	40
No	60	50

Figure 4: Example Contingency Table

About the contingency table.

- The cells are the possible outcomes.
- The cell probabilities are the multinomial parameters.

Suppose n_{ij} observations in the ij-th cell, $i = 1, 2$; $j = 1, 2$.

	Factor Y	
Factor X	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

Figure 5: Notations in Contingency Table

2.3.2 Hypothesis Test

Let π_{ij} denote the true unknown joint probability of falling into ij-th cell, $i = 1, 2$, $j = 1, 2$. We have

- $\pi_{11} = \mathbf{P}(\text{Yes and Yes})$.
- $\pi_{12} = \mathbf{P}(\text{Yes and No})$.
- $\pi_{21} = \mathbf{P}(\text{No and Yes})$.
- $\pi_{22} = \mathbf{P}(\text{No and No})$.

The cell probabilities can be represented as follows.

	Factor Y		
Factor X	Yes	No	Total
Yes	π_{11}	π_{12}	$\pi_1 = \pi_{11} + \pi_{12}$
No	π_{21}	π_{22}	$\pi_2 = \pi_{12} + \pi_{22}$
Total	$\pi_{.1} = \pi_{11} + \pi_{21}$	$\pi_{.2} = \pi_{12} + \pi_{22}$	1

Figure 6: Parameters in Contingency Table

We would like to set the null hypothesis as that all joint probabilities are equal to the product of their marginal probabilities. So we have

$$H_0 : \pi_{ij} = \pi_i \pi_{.j}, i = 1, 2, j = 1, 2$$

Under H_0, we have

	Belief in Afterlife	
Belief in Aliens	Yes	No
Yes	$\pi_{11} = \pi_1 \pi_{.1}$	
No		

Figure 7: Contingency Example under H_0

The expected value of n_{11} is $n\pi_{11}$, so under the null hypothesis, $n\pi_{11} = n\pi_1\pi_{\cdot 1}$.

Moreover, π_1 can be estimated by $\frac{n_{11}+n_{12}}{n}$ and $\pi_{\cdot 1}$ can be estimated by $\frac{n_{11}+n_{21}}{n}$.

Based on Fig 5, under H_0, the estimates of the expected values are

$$
\begin{aligned}
E_{11} &= \frac{(n_{11}+n_{12})(n_{11}+n_{21})}{n};\\
E_{12} &= \frac{(n_{11}+n_{12})(n_{12}+n_{22})}{n};\\
E_{21} &= \frac{(n_{11}+n_{21})(n_{21}+n_{21})}{n};\\
E_{22} &= \frac{(n_{12}+n_{22})(n_{21}+n_{22})}{n}.
\end{aligned}
$$

The χ^2 test statistics is as follows

$$
S = \frac{(n_{11}-E_{11})^2}{E_{11}} + \frac{(n_{12}-E_{12})^2}{E_{12}} + \frac{(n_{21}-E_{21})^2}{E_{21}} + \frac{(n_{22}-E_{22})^2}{E_{22}}
$$

We reject H_0 if $S \geq \chi^2_{\alpha,1}$.

2.3.3 Action in R

```
n11=50; n12=40; n21=60; n22=50;
xm=matrix(c(n11,n12,n21,n22),2,2)
xm
# [,1] [,2]
# [1,] 50 60
# [2,] 40 50
chisq.test(xm,correct=F)
# Pearson's Chi-squared test
#
```

```
# data: xm
# X-squared = 0.020406, df = 1, p-value = 0.8864

n1=n11+n12
n2=n21+n22
prop.test(c(n11,n21),c(n1,n2),correct=F)
# 2-sample test for equality of proportions without continuity correction
#
# data: c(n11, n21) out of c(n1, n2)
# X-squared = 0.020406, df = 1, p-value = 0.8864
# alternative hypothesis: two.sided
# 95 percent confidence interval:
# -0.1284537 0.1486558
# sample estimates:
# prop 1 prop 2
# 0.5555556 0.5454545

chitest=chisq.test(xm,correct=F)
names(chitest)
# [1] "statistic" "parameter" "p.value" "method" "data.name"
# "observed" "expected" "residuals"
# [9] "stdres"
chitest$observed
# [,1] [,2]
# [1,] 50 60
# [2,] 40 50
chitest$expected
# [,1] [,2]
# [1,] 49.5 60.5
# [2,] 40.5 49.5
```