Applied Statistical Methods

Analysis of Categorical Data

Xuemao Zhang
East Stroudsburg University

March 15, 2023

Outline

- One Proportion
  - Central Limit Theorem
  - Estimations and Inferences
- One-way Frequency Table
  - Multinomial Experiment
  - One-way Chi-Square Test (Goodness-of-Fit Test)
- Contingency Tables
  - Chi-Square Contingency Test
  - Fisher's Exact Test
  - McNemar's Test
  - Cochran's Q Test

One Proportion

- The C.L.T. can also be applied to the sample proportions
- If Y is a binomial(n, p) random variable, then we can write $Y = \sum_{i=1}^{n} X_i$ where

$$X_i = \begin{cases} 1, & \text{if the ith trial is a success;} \\ 0, & \text{Otherwise.} \end{cases}$$

  By the the central limit theorem, for large n, the sample proportion $\hat{p} = \frac{Y}{n} = \overline{X}$ is approximately normal with mean p and variance $p(1-p)/n$. That is, approximately

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right).$$

- Large sample size requirements: $np \geq 10$ and $n(1-p) \geq 10$.

One Proportion

- A $100(1-\alpha)\%$ confidence interval of p is

$$\hat{p} \pm Z_{\alpha/2} SE(\hat{p}),$$

where $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- Hypothesis testing:

$$H_0 : p = p_0$$

$$H_a : \begin{cases} p > p_0, & \text{upper-tail alternative;} \\ p < p_0, & \text{lower-tail alternative;} \\ p \neq p_0, & \text{two-tailed alternative.} \end{cases}$$

Test Statistic: $Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

$$p - value = \begin{cases} P(Z \geq Z_0), & \text{upper-tail area;} \\ P(Z \leq Z_0), & \text{lower-tail area;} \\ 2P(Z \geq |Z_0|), & \text{twice of smaller tail area.} \end{cases}$$

One Proportion

- Example: Given a random sample of M&Ms of size 100, 19 of them are green. Use the sample information to test the claim of Mars, Inc., that 16% of its plain M&M candies are green.
- Function `statsmodels.stats.proportion.proportions_ztest()`
  - https://www.statsmodels.org/dev/generated/statsmodels.stats.proportion.proportions_ztest.html

```
import numpy as np
from statsmodels.stats.proportion import proportions_ztest
stat, pval = proportions_ztest(count=19, nobs=100, value=0.16,
alternative = "two-sided")
print(stat,pval)
```

```
## 0.7647191129018724 0.4444388225040353
```

```
# R code:
prop.test(19,100,p=0.16,alternative = "two.sided",correct = F)
```

Multinomial Experiment

Many experiments result in measurements that are qualitative or categorical rather than quantitiative. The following characteristics, which define a multinomial experiment:

- The experiment consists of **n identical trials**. (binomial)
- Each trial results in **one of k** categories.
- The probability that the outcome falls into a particular category i on a single trial is p_i and **remains constant** from trial to trial. The sum of all k probabilities,
$$p_1 + p_2 + \cdots + p_k = 1.$$
- The trials are **independent**.
- We are interested in the number of outcomes in each category, $O_1, O_2, \ldots, O_k$ with $O_1 + O_2 + \cdots + O_k = n$.

Multinomial Experiment

In the multinomial experiment, we make inferences about all the probabilities, $p_1, p_2, p_3, \ldots, p_k$.

It can be shown that the expected number of outcomes resulting in category i is

$$\boxed{E(O_i) = np_i, \quad i = 1, 2, \ldots, k.}$$

Suppose that we hypothesize values for $p_1, p_2, \ldots, p_k$ and calculate the expected value for each cell. Certainly if our hypothesis is true, the cell counts ni should not deviate greatly from their expected values np_i for $i = 1, 2, \ldots, k$. Hence, it would seem intuitively reasonable to use a test statistic involving the k deviations,

$$O_i - E(O_i) = O_i - np_i, \quad i = 1, 2, \ldots, k.$$

One-way Chi-Square Test

In 1900 Karl Pearson proposed the following test statistic

$$X^2 = \sum_{i=1}^{k} \frac{[O_i - E(O_i)]^2}{E(O_i)} = \sum_{i=1}^{k} \frac{[O_i - np_i]^2}{np_i}.$$

It can be shown that when n is large, X^2 has an approximate chi-square (χ^2) probability distribution. We can easily demonstrate this result for the case $k = 2$ as follows.

One-way Chi-Square Test

If $k = 2$, then $O_2 = n - O_1$ and $p_1 + p_2 = 1$. Thus,

$$
\begin{aligned}
X^2 = & \sum_{i=1}^{k} \frac{[O_i - E(O_i)]^2}{E(O_i)} = \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 - np_2)^2}{np_2} \\
= & \frac{(O_1 - np_1)^2}{np_1} + \frac{[(n - O_1) - n(1 - p_1)]^2}{n(1 - p_1)} \\
= & \frac{(O_1 - np_1)^2}{np_1} + \frac{(-O_1 + np_1)^2}{n(1 - p_1)} \\
= & (O_1 - np_1)^2 \left(\frac{1}{np_1} + \frac{1}{n(1 - p_1)} \right) \\
= & \frac{(O_1 - np_1)^2}{np_1(1 - p_1)}
\end{aligned}
$$

One-way Chi-Square Test

By C.L.T., for large n,

$$\frac{O_1 - np_1}{\sqrt{np_1(1-p_1)}}$$

has approximately a standard normal distribution. Since the square of a standard normal random variable has a χ^2 distribution, for $k = 2$ and large n, X^2 has an approximate χ^2 distribution with 1 df (degree of freedom).

One-way Chi-Square Test

- Sample size requirement: Experience has shown that the cell counts n_i should not be too small if the χ^2 distribution is to provide an adequate approximation to the distribution of X^2. As a rule of thumb, we will require that **all expected cell counts are at least five**, although Cochran (1952) has noted that this value can be as low as one for some situations.

- Determine df: The **principle** to determine the df is: *the appropriate number of degrees of freedom will equal the number of cells, k, less 1 df for each independent linear restriction placed on the cell probabilities.*

One-way Chi-Square Test

Suppose the categorical variable has c categories, and that the population proportion in category i is p_i. To test

$$\begin{aligned} H_0: \quad p_i &= p_i^{(0)}, \ i = 1, 2, \ldots, k \\ H_a: \quad p_i &\neq p_i^{(0)} \text{ for at least one } i \end{aligned}$$

for pre-specified values $p_i^{(0)}, i = 1, 2, \ldots, k$, use the **Pearson χ^2 statistic**

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - np_i^{(0)})^2}{np_i^{(0)}},$$

where O_i is the observed frequency in category i, and n is the total number of observations.

One-way Chi-Square Test

Note that for each category the Pearson statistic computes **(observed-expected)^2/expected** (noting that we assume H_0 true and under this assumption, the expected number in category i is $np_i^{(0)}$) and sums over all categories.

When H_0 is true, the differences observed-expected for all cells will be small, but large when H_0 is false. We reject H_0 only if X^2 is **large**.

If there is sufficient data (Guideline: The expected number in each category is at least 5), then under H_0, $X^2 \sim \chi^2_{\mathbf{k-1}}$. Therefore, if x^{2*} is the observed value of X^2 calculated from the data, the p-value of the test is

$$P(\chi^2_{k-1} \geq x^{2*}).$$

One-way Chi-Square Test

- Example: A random sample of 100 weights of Californians is obtained, and the last digits of those weights are summarized in the following table. Test the claim that the sample is from a population of weights in which the last digits do not occur with the same frequency. The raw data are of the form

Last Digit	0	1	2	3	4	5	6	7	8	9
Frequency	46	1	2	3	3	30	4	0	8	3

- Function `scipy.stats.chisquare()`
  - https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html

```
from scipy.stats import chisquare
x=[46,1,2,3,3,30,4,0,8,3] #p0=[1/10]*10
stat, pval=chisquare(x)
print(stat, pval)
```

```
## 212.8 6.8250225439989 85e-41
```

```
# R code
x=c(46,1,2,3,3,30,4,0,8,3);
chisq.test(x, p=rep(1/10,10));
```

Contingency Table

Analysis of categorical data is based on counts, proportions or percentages of data that fall into the various categories defined by the variables.

Suppose a population is partitioned into rc categories, determined by r levels of variable 1 and c levels of variable 2. The population proportion for level i of variable 1 and level j of variable 2 is p_{ij}. This information can be displayed in the following $r \times c$ table:

Two-Way Table of Proportions

	Column				Marginals
row	1	2	...	c	
1	p_{11}	p_{12}	...	p_{1c}	$p_{1\cdot}$
2	p_{21}	p_{22}	...	p_{2c}	$p_{2\cdot}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	p_{r1}	p_{r2}	...	p_{rc}	$p_{r\cdot}$
Marginals	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot c}$	1

Contingency Table

Two-Way Table of Counts

	Column				Marginals
row	1	2	...	c	
1	O_{11}	O_{12}	...	O_{1c}	$R_{1.}$
2	O_{21}	O_{22}	...	O_{2c}	$R_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	O_{r1}	O_{r2}	...	O_{rc}	$R_{r.}$
Marginals	$C_{.1}$	$C_{.2}$	...	$C_{.c}$	n

We want to test

H_0 : row and column variables are independent

H_a : row and column variables are not independent.

Chi-Square Contingency Test

To do so, we select a random sample of size n from the population. Suppose the table of observed frequencies is

	Column				Totals
row	1	2	...	c	
1	O_{11}	O_{12}	...	O_{1c}	$R_{1.}$
2	O_{21}	O_{22}	...	O_{2c}	$R_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	O_{r1}	O_{r2}	...	O_{rc}	$R_{r.}$
Totals	$C_{.1}$	$C_{.2}$	...	$C_{.c}$	n

It can be shown that under H_0 the expected cell frequency for the ij cell is given by

$$\begin{aligned} E_{ij} &= \frac{\text{row i total} \times \text{column j total}}{\text{sample size}} \\ &= \frac{R_{i.} C_{.j}}{n} = n\hat{p}_{i.}\hat{p}_{.j}, \end{aligned}$$

where $\hat{p}_{i.} = R_{i.}/n$ and $\hat{p}_{.j} = C_{.j}/n$.

Chi-Square Contingency Test

- Proof

Chi-Square Contingency Test

To measure the deviations of the observed frequencies from the expected frequencies under the assumption of independence, we construct the Pearson χ^2 statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

If H_0 is true, X^2 has (approximately) a $\chi^2_{(\mathbf{r}-\mathbf{1})(\mathbf{c}-\mathbf{1})}$ distribution.

(Note that for the approximation to be valid, we require that $E_{ij} \geq 5$).

Chi-Square Contingency Test

Example A survey was conducted to evaluate the effectiveness of a new flu vaccine that had been administered in a small community. The vaccine was provided free of charge in a two-shot sequence over a period of 2 weeks to those wishing to avail themselves of it. Some people received the two-shot sequence, some appeared only for the first shot, and the others received neither.

A survey of 1000 local inhabitants in the following spring provided the information shown in the following table. Do the data present sufficient evidence to indicate a dependence between the two classifications - vaccine category and occurrence or nonoccurrence of flu?

Status	No Vaccine	One Shot	Two Shots	Total
Flu	24 (14.4)	9 (5.0)	13 (26.6)	46
No flu	289 (298.6)	100 (104.0)	565 (551.4)	954
Total	313	109	578	1000

Chi-Square Contingency Test

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html

```
import numpy
from scipy import stats
obs=numpy.array([[24,9,13],[289,100,565]])
g, p, dof, expctd=stats.chi2_contingency(obs)
print(g, p, dof, expctd)
```

```
## 17.31297056699376 0.00017399477577242703 2 [[ 14.398   5.014  26.588]
##  [298.602 103.986 551.412]]
```

```
# R code:
No_Vaccine=c(24, 289)
One_Shot=c(9, 100)
Two_Shots=c(13, 565)
Dataset = as.data.frame(cbind(No_Vaccine,One_Shot, Two_Shots))
row.names(Dataset)=c("Flu", "No_Flu")
test1=chisq.test(Dataset, correct=FALSE)
test1
test1$observed
test1$expected
```

Chi-Square Contingency Test - Test of Homogeneity

Sometimes researchers design an experiment so that the number of experimental units falling in one set of categories is **fixed in advance**.

Each of the c columns (or r rows) whose totals have been fixed in advance is actually a single multinomial experiment. For example: An experimenter selects 900 patients who have been treated for flu prevention. She selects 300 from each of three types - no vaccine, one shot, and two shots.

Status	No Vaccine	One Shot	Two Shots	Total
Flu				r_1
No flu				r_2
Total	300	300	300	900

Chi-Square Contingency Test - Test of Homogeneity

Without loss of generality, suppose the {**column totals are fixed in advance**}. **That is, there are** c **populations.**

	Column				
row	**1**	**2**	**...**	c	**Total**
1	O_{11}	O_{12}	**...**	O_{1c}	R_1
2	O_{21}	O_{22}	**...**	O_{2c}	R_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	O_{r1}	O_{r2}	**...**	R_{rc}	R_r
Totals	C_1	C_2	**...**	C_c	$n = C_1 + \cdots + C_c$

Chi-Square Contingency Test - Test of Homogeneity

We have c multinomial populations with probabilities

Two-Way Table of Proportions

	Column				
row	1	2	...	c	
1	p_{11}	p_{12}	...	p_{1c}	
2	p_{21}	p_{22}	...	p_{2c}	
.	.	.		.	
.	.	.		.	
.	.	.		.	
r	p_{r1}	p_{r2}	...	p_{rc}	
	1	1	...	1	

• In a test of homogeneity, we test the claim that different populations have the same proportions of some characteristics. The chi-square test of Homogeneity is **equivalent** to a test of the equality of c multinomial populations (suppose the columns are fixed in advance).

Chi-Square Contingency Test - Test of Homogeneity

Distinguish between a Test of Homogeneity and a Test for Independence:

- In a typical test of independence, sample subjects are randomly selected from **one population** and values of two different variables are observed.
- In a test of homogeneity, subjects are randomly selected from the c **different populations** separately.

Now, we are testing:

$$H_0 : p_{i1} = p_{i2} = \cdots = p_{ic} \qquad i = 1, 2, \ldots, r$$
$$H_A : \text{ Not all } p_{i1}, p_{i2}, \ldots, p_{ic} \text{ are equal for some } i \qquad i = 1, 2, \ldots, r$$

Remark. The testing procedure is the same as the test of independence except the explanations are different:

Test Statistic:

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

Chi-Square Contingency Test - Test of Homogeneity

where

$$E_{ij} = \frac{\text{row i total} \times \text{column j total}}{\text{sample size}} = \frac{R_i . C_j}{n}.$$

- Under H_0, the test statistic has an approximate Chi-square distribution with $df = (r-1)(c-1)$.
- The Chi-square test is upper-tailed. That is, H_0 should be rejected only if the calculated X^2 is large.

Example. A survey of voter sentiment was conducted in four midcity political wards to compare the fraction of voters favoring candidate A. Random samples of 200 voters were polled in each of the four wards, with results as shown in the follwing Table. Do the data present sufficient evidence to indicate that the fractions of voters favoring candidate A differ in the four wards?

Chi-Square Contingency Test - Test of Homogeneity

	Ward			
Opinion	1	2	3	4
Favor A	76	53	59	48
Do not favor A	124	147	141	152
Total	200	200	200	200

```
import numpy
from scipy import stats
obs=numpy.array([[76,53,59,48],[124,147,141,152]])
g, p, dof, expctd=stats.chi2_contingency(obs)
print(g, p, dof, expctd)
```

```
## 10.722442601274192 0.013325431052795 21 3 [[ 59.  59.  59.  59.]
##  [141. 141. 141. 141.]]
```

```
# R code
Ward1=c(76, 124)
Ward2=c(53, 147)
Ward3=c(59, 141)
Ward4=c(48, 152)
dataset = as.data.frame(cbind(Ward1,Ward2, Ward3,Ward4))
rownames(dataset) = c('Opinion1', 'Opinion2')
test2=chisq.test(dataset, correct=FALSE)
test2
test2$observed
```

Fisher's Exact Test

- Fisher's exact test is for 2×2 contingency tables.
- Fisher's exact test should be used when 20% or more cells has expected counts less than 5.
- Fisher exact test is based on the hypergeometric distribution (https://en.wikipedia.org/wiki/Fisher%27s_exact_test).
- Fisher exact test returns the exact p-value.
- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.fisher_exact.html

```
import numpy
from scipy import stats
table = np.array([[6, 2], [1, 4]])
oddsr, p=stats.fisher_exact(table)
print(oddsr, p)
```

```
## 12.0 0.10256410256410256
```

```
# R function fisher.test()
```

McNemar's Test

- McNemar's test is for 2×2 contingency tables.
- It is a test of association which examines the relationship that exists among the cells of the table. The McNemar's test examines the difference between the proportions that derive from the marginal sums of the table (see below): $p_A = \frac{a+b}{N}$ and $p_B = \frac{a+c}{N}$.

	B		
A	0	1	Total
0	a	b	$a+b$
1	c	d	$c+d$
Total	$a+c$	$b+d$	$N = a+b+c+d$

- The McNemar's test checks if p_A and p_B are significantly different.

McNemar's Test

- Suppose 0 means failure and 1 means success. Also let B be the test and A be the control
- In order to test if the treatment is helpful, we use only the number discordant pairs of twins, b and c, since the other pairs of twins tell us nothing about whether the treatment is helpful or not. McNemar's test is

$$Q = Q(b, c) = \frac{(b - c)^2}{b + c}$$

which for large samples is distributed like a chi-squared distribution with 1 degree of freedom. A closer approximation to the chi-squared distributi0n uses a continuity correction:

$$Q_C = Q_C(b, c) = \frac{(|b - c| - 1)^2}{b + c}$$

McNemar's Test

- McNemar's test can be used for example in studies in which patients serve as their own control, or in studies with "before and after" design.
- Example: A researcher attempts to determine if a drug has an effect on a particular disease. The test requires the same subjects to be included in the before-and-after measurements (matched pairs).
  - The null hypothesis of "marginal homogeneity" means there was no effect of the treatment.

	After:present	After:absent	Total
Before:present	101	121	222
Before:absent	59	33	92
Total	160	154	314

McNemar's Test

- https://www.statsmodels.org/dev/generated/statsmodels.sandbox.stats.runs.mcnemar.html

```
from statsmodels.sandbox.stats.runs import mcnemar
obs=[[101,121], [59,33]]
chi2,p=mcnemar(obs,exact=False, correction=True)
print(chi2,p)
```

```
## 20.67222222222222 5.450094825427117e-06
```

```
# R code
After_present=c(101, 59)
After_absent=c(121, 33)
data0 = as.matrix(cbind(After_present, After_absent))
#the function requires the input be a matrix
mcnemar.test(data0, correct = TRUE)
```

Cochran's Q Test

- In the analysis of **two-way randomized block designs** where the response variable is **binary (coded as 0 and 1)**, Cochran's Q test is a non-parametric statistical test to verify whether the treatments have identical effects (https://en.wikipedia.org/wiki/Cochran%27s_Q_test).
- Null hypothesis (H_0): the treatments are equally effective. Alternative hypothesis (H_a): there is a difference in effectiveness between treatments.

Cochran's Q Test

- Example: Twelve subjects are asked to perform three tasks. The outcome of each task is success (1) or failure(0).

Subject	Task1	Task2	Task3
1	0	1	0
2	1	1	0
3	1	1	1
4	0	0	0
5	1	0	0
6	0	1	1
7	0	0	0
8	1	1	0
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0

Cochran's Q Test

- https://www.statsmodels.org/dev/generated/statsmodels.sandbox.stats.runs.cochrans_q.html

```
import numpy as np
from statsmodels.stats.contingency_tables import cochrans_q
data=np.array([[0,1,1,0,1,0,0,1,0,0,0,0],
[1,1,1,0,0,1,0,1,1,1,1,1],
[0,0,1,0,0,1,0,0,0,0,0,0]])
data1=np.transpose(data)
print( cochrans_q(data1) )
```

```
## df          2
## pvalue      0.01312372873694097
## statistic   8.666666666666666
```

```
# R code
Task1=c(0,1,1,0,1,0,0,1,0,0,0,0)
Task2=c(1,1,1,0,0,1,0,1,1,1,1,1)
Task3=c(0,0,1,0,0,1,0,0,0,0,0,0)
data1 = as.matrix(cbind(Task1, Task2, Task3))
library(nonpar)
cochrans.q(data1, alpha=0.05)
```

License

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.