Goodness-of-fit Chi-Squared Tests
Categorical Data inference

In this section we will study catergorical data. Consider a experiment, which has a finite sample space $S=\{s_1, s_2, \ldots, s_k\}$ associated with probability p_i . If we repeat n experiments, then each outcome s_i occurs N_i times with $N_i \geqslant 0$ and $\sum_{i=1}^{k} N_i = n$. Random variables $N_1, N_2, \ldots, N_k$ follows so-call multinomial distribution, and its observed values are denoted as $n_1, n_2, \ldots, n_k$.
We then interested in making inference for p_i's, especially hypothesis testing.

Note If k=2, this reduces to binomial distribution problem with unknown population proportion.

Testing for H_0: $p_i = p_{i0}$ for all i=1,..,k

We first condition the following hypothesis testing problem where the null probability values are completely specified. That is

$$H_0: p_1 = p_{10} \text{ and } \ldots \text{ and } p_k = p_{k0} \text{ vs } H_1: \text{otherwise}$$

where all p_{i0} are all given constants.

Under null hypothesis,

$$E(N_i) = np_{i0}$$

This imples that larger deviation ($n_i - E(N_i) = n_i - np_{i0}$) means that more contradictory between data and null hypothesis. Thus we construct a test statistic that combines all $n_i - np_{i0}$.

Theorem

Under null hypothesis, the statistic $\sum_{i=1}^{k} \frac{(N_i - np_{i0})^2}{np_{i0}}$ has a approximated distribution chi-squared dsitribution with df k-1, if all np_{i0}>5.

Therefore,

Test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{k} \frac{(\text{observed-expected})^2}{\text{expected}}$$

where "expected" means *expected number under null hypothesis.*

Rejection region

$\chi^2 > \chi^2_{\alpha, k-1}$, and p-value =

Note

When k=2, our CLT-based test statistic is $Z=\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$, with rejection region $|Z|>Z_{\alpha/2}$.

This is equivalent to test statistic

$$Z^2=\frac{(\hat{p}-p_0)^2}{p_0(1-p_0)/n}=\frac{(n\hat{p}-np_0)^2}{np_0(1-p_0)}=\frac{(n\hat{p}-np_0)^2}{np_0}+\frac{(n(1-\hat{p})-n(1-p_0))^2}{n(1-p_0)}=\sum_{i=1}^{2}\frac{(\text{observed-expected})^2}{\text{expected}}$$

with rejection region $Z^2>Z^2_{\alpha/2}=\chi^2_{\alpha,1}$.
So, this is consist to the above test statistic.

Example:
If we focus on two different characteristics of an organism, each controlled by a single gene, and cross a pure strain having genotype AABB with a pure strain having genotype aabb (capital letters denoting dominant alleles and small letters recessive alleles), the resulting genotype will be AaBb. If these first-generation organisms are then crossed among themselves (a dihybrid cross), there will be four phenotypes depending on whether a dominant allele of either type is present. Mendel's laws of inheritance imply that these four phenotypes should have probabilities 9/16 , 3/16 , 3/16 and 1/16 of arising in any given dihybrid cross.
A scientific reports the following data on phenotypes from a dihybrid cross of tall cut-leaf tomatoes with dwarf potato-leaf tomatoes.

Tall, cut leaf	Tall, potato leaf	Dwarf, cut leaf	Dwarf, potato leaf
926	288	293	104

Please test whether Mandel's law holds.

Example:
$X_1, \ldots, X_n$ are independent Binomial(4, p) random variables. Propose a chi-square test for H_0: $p=p_0$ vs H_0: p is not p_0.

In the above example, do we miss some information?

Example:
Test a underlying continuous distribution.

Testing for H_0: $p_i = p_i(\theta)$ for some θ and all i=1,..,k

Denote $\vec{p}$ = $(p_1,\ldots,p_k)$ be a k-dimensional vector of parameter
Simple hypothesis: H_0: $\vec{p}=\vec{p}_0$ where $\vec{p}_0$ is a single k-dimensional vector.
Composite hypothesis: H_0: $\vec{p}\in\Theta$ where $\Theta\subset R^k$ is set with multiple elements.

In this section, we consider $\Theta=\{\vec{p}_0(\vec{\theta}):\vec{\theta}\in R^m\}$ where $\vec{p}_0(\cdot)=(p_{01}(\cdot),\ldots,p_{0k}(\cdot))$ is a known vector-value function with a m-dimensional argument. Therefore, the null hypothesis $H_0:\vec{p}\in\Theta$ is equivalent to

$$H_0:\vec{p}=\vec{p}_0(\vec{\theta}) \text{ for some } \vec{\theta}$$

Example
$X_1, \ldots, X_n$ are independent random variables which only take value 0,1,2,3,4. We want to test whether they follows a binomial distribution.

Testing procedure:

1. Assuming null hypothesis is true, we estimate $\vec{\theta}$ by using MLE of the data $\hat{\vec{\theta}}$.
2. Test statistic:
$$\chi^2=\sum_{i=1}^{k}\frac{(N_i-np_{0i}(\hat{\vec{\theta}}))^2}{np_{0i}(\hat{\vec{\theta}})}$$
3. Under null hypothesis, $\chi^2\sim\chi^2_{k-1-m}$ if $np_{0i}(\hat{\vec{\theta}})>5$ for all i=1, … ,k.
4. Reject null hypothesis if the observed test statistic is larger than $\chi^2_{\alpha,k-1-m}$

Example

X_1, …, X_n are independent random variables which only take value 0,1,2,3,4. We want to test whether they follows a binomial distribution.

1. If indeed all X's follows Bin(n,p). What is the MLE $\hat{p}$.

2. The estimated null probability is
$p_{01}(\hat{p})=$
$p_{02}(\hat{p})=$
$p_{03}(\hat{p})=$
$p_{04}(\hat{p})=$
$p_{05}(\hat{p})=$

3.
$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_{0i}(\hat{p}))^2}{np_{0i}(\hat{p})}$$

4. Conclusion

Testing for contingency table

Example: Researchers wanted to test the theory that *women who went to work shortly after giving birth were more likely to experience postpartum depression compared to those who stayed home.*

A random sample of women giving birth at a Dallas hospital were queried six months after giving birth

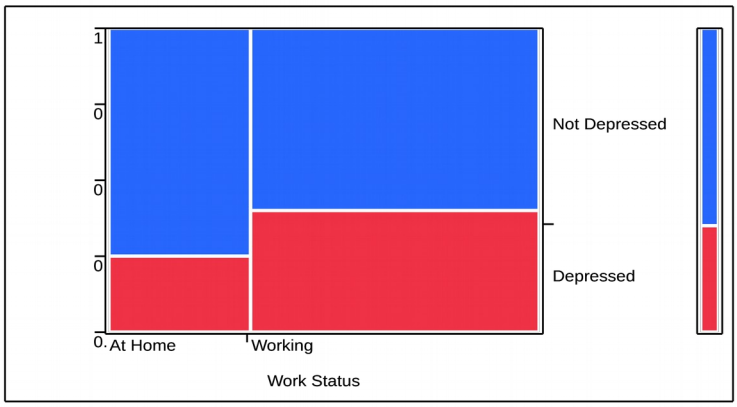
to their first child. The researchers recorded whether or not the woman worked outside the home and whether or not she experienced postpartum depression.

Descriptive Statistics:

Contingency Table

Work Status By Mental State

Count Row % Expected	Depressed	Not Depressed	
At Home	17	50	67
Working	55	81	136
	72	131	203



Contingency table:

Case 1:

- The rows are the different n populations– one row for each population. These populations are defined by the value of a categorical variable.
- In each population, we make random samples **separately**. *That is, the row sum is not proportional to the population*.
- The columns are divided up by another variable value – one column for each response variable value.
- The number in a cell (box) is the number of subjects in that row population taking that column's response value.

Case 2:

- There is a grand population associated with 2 random variables.
- We make a random sample from this population.
- Each cell gives the number of subjects in the sample that has the corresponding values for the 2 random variable.

Assume there is R rows and C columns, the count is denoted by n_{ij}. And rum is denoted by $n_{i\cdot}$, column sum is denoted by $n_{\cdot j}$.

Null Hypotheses

Case 1:

The populatoin proportion of each value of the column variables are the same across all population.

For the i^{th} population, its population proportion for column variable is denotes as p_{i1}, p_{i2},…, p_{ic}

$$H_0: p_{1j} = p_{2j} = ... = p_{Rj} \quad \text{for all j=1,…,C.}$$

Case 2:

In the population there is no association between 2 variables (independence).

The population proportion of (i,j) cell is denoted by p_{ij},

$$H_0: p_{ij} = p_{i\cdot} \cdot p_{j} \quad \text{for all i=1,…,R and j=1,…,C for some } p_i\text{'s and } p_j\text{'s.}$$

Derive test statistics under case 1:

1. If Null hypothesis is true, what is the MLE $\hat{p}_{ij}$.

$$\hat{p}_{ij} = n_{\cdot j}/n$$

2. The estimated expected value for (I,j) cell is

$$n_{i\cdot}\,\hat{p}_{ij} = n_{i\cdot}\,n_{\cdot j}/n$$

3. Test statistics is

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(n_{ij} - n_{i\cdot}\,n_{\cdot j}/n\right)^2}{n_{i\cdot}\,n_{\cdot j}/n}$$

4. Degree of freedom is R(C-1)-1-C=(R-1)(C-1)
5. Rejection region is:

Derive test statistics under case 2:

1. If Null hypothesis is true, what is the MLE $\hat{p}_i$ and $\hat{p}_j$.

$$\hat{p}_i = n_{i\cdot}/n \text{ and } \hat{p}_j = n_{\cdot j}/n$$

2. The estimated expected value for (I,j) cell is

$$n_{i\cdot}\,\hat{p}_i\,\hat{p}_j = n_{i\cdot}\,n_{\cdot j}/n$$

3. Test statistics is

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{\left(n_{ij} - n_{i\cdot}\,n_{\cdot j}/n\right)^2}{n_{i\cdot}\,n_{\cdot j}/n}$$

4. Degree of freedom is RC-1-(R-1)-(C-1)=(R-1)(C-1)
5. Rejection region is

Same test statistic and rejection region.

Example cont:

For our data set the TS is calculated as

TS = $\frac{(17-23.7635)^2}{23.7635} + \frac{(50-43.2365)^2}{43.2365} +$
$\frac{(55-48.2365)^2}{48.2365} + \frac{(81-87.7635)^2}{87.7635}$ = 4.4527

Note: Large sample sizes requirement - Expected number of observations in each cell ≥ 5.

Count Row % Expected	Depressed	Not Depressed	
At Home	17 25.37 23.7635	50 74.63 43.2365	67
Working	55 40.44 48.2365	81 59.56 87.7635	136
	72	131	203