Packet 7: More Testing

Chap 9.1 Chi-Square Goodness-of-Fit Test

The frequency table and the multinomial distribution:

Sample units are classified into K mutually exclusive categories, the number of units falling into each category is recorded.

The frequency table is a "One way" table in which units are classified according to a single categorical variable.

E.g., Eye color: Brown, Blue, Black, Green, and others. (unordered categorical variable or nominal variable).

E.g., Attitude toward war: strongly agree, agree, disagree, strongly disagree. (ordered but no numerical scores).

E.g., number of children in a family: 0,1,2, ..., 22. (ordered with numerical values).

General setting:

Suppose a random experiment has K possible outcomes, say $A_1, A_2, \cdots, A_K$.

Let $p_i = P(A_i)$, and thus $\sum_i p_i = 1$.

Repeat experiments n times independently. Let X_i be the number of A_i in n trials.

Assumptions:

$X = (X_1, X_2, \cdots, X_K) \sim$ multinomial distribution(n, p), where n is often known and $p = (p_1, p_2, \cdots, p_k)$. The probability mass function is

$$f(x_1, x_2, \cdots, x_k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

The critical assumptions:

- N trials are independent !!!
- The parameter p remains constant from trial to trial.

The most common violation occurs when clustering is presented in the data.

E.g., Suppose eye color samples were not collected from unrelated individuals , but from multiple families. Persons within a family are more likely to have the same color than persons from different families.

Pearson (Chi-square) goodness of fit

H_0: $p_1 = p_{10}, p_2 = p_{20}, \cdots, p_K = p_{K0}$, pre defined values

H_1: at least one equation does not hold.

Special case: Let $K = 2$, $X_1 \sim Bin(n, p_1)$, $X_2 \sim Bin(n, p_2)$. $X_1 + X_2 = n$ $\quad P_1 + P_2 = 1$

X_1 and X_2 are not indepedent, $X_2 = n - X_1$ is fully determined after knowing X_1

$$H_0 \; P_1 = P_{10} \quad \text{v.s.} \quad H_1 \; P_1 \neq P_{10}$$

$$Z = \frac{\hat{P}_1 - P_{10}}{\sqrt{P_{10}(1-P_{10})/n}} \;\dot{\sim}\; N(0,1)$$

where $\hat{P}_1 = X_1/n$

$$Z^2 = \frac{(\hat{P}_1 - P_{10})^2 \times n^2}{P_{10}(1-P_{10})/n \times n^2} \;\dot{\sim}\; \chi^2(1)$$

$$= \frac{(X_1 - nP_{10})^2}{n P_{10}(1-P_{10})}$$

$$= \frac{(X_1 - nP_{10})^2}{nP_{10}} + \frac{(X_1 - nP_{10})^2}{n(1-P_{10})}$$

$$= \frac{(X_1 - nP_{10})^2}{nP_{10}} + \frac{(X_2 - nP_{20})^2}{nP_{20}}$$

$$= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

chi-square test statistic

① $$\frac{1}{P_{10}(1-P_{10})} = \frac{1-P_{10}}{P_{10}(1-P_{10})} + \frac{P_{10}}{P_{10}(1-P_{10})} = \frac{1}{P_{10}} + \frac{1}{1-P_{10}}$$

② $1 - P_{10} = P_{20}$

$$X_1 + X_2 = n = nP_{10} + nP_{20}$$

$$(X_1 - nP_{10})^2 = (nP_{20} - X_2)^2$$

If H_0 is wrong, then Z^2 will be large.

For $i = 1, 2$

X_i is the observed count in category i, denoted by O_i

nP_{i0} is the expected count under H_0 [2], denoted by E_i

$P_1 + P_2 + \cdots P_K = 1$

$K-1$ d.f.

Extend to k categories:

Critical Region

$Q \geq \chi^2_\alpha (K-1)$

under H_0

$$Q = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(K-1)}.$$

qchisq$(1-\alpha, K-1)$

If O_i is far away from E_i for some categories then test statistic Q is large, reject H_0.

C.L.T.

$Q \sim \chi^2$ if n is large enough to have $E_j = np_j < 5$ for no more than 20% of the cells in the table. None of E_j should fall below 1.

If the above condition is not satisfied, we often combine cells until all E_j are large enough.

If n is not large, instead of assuming asymptotic distributions, play with exact numbers in the table. E.g. Fisher's exact test. Not cover in 415

$K=4$

Example 1: A bag of candies have 4 colors. Test if the 4 colors are in equal proportions at $\alpha = 0.05$.

$H_0 : p1 = p2 = p3 = p4$ $= 0.25$ v.s. H_1: not all equal.

$n = 224$, observe $X_1 = 42$, $X_2 = 64$, $X_3 = 53$, $X_4 = 65$.

$O_1 \quad O_2 \quad O_3 \quad O_4$

$E_1 = E_2 = E_3 = E_4 = n \times 0.25 = 224 \times 0.25 = 56$

$$Q = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(42-56)^2}{56} + \frac{(64-56)^2}{56} + \frac{(53-56)^2}{56} + \frac{(65-56)^2}{56}$$

$$= 6.25$$

Critical Region $\quad Q \geq \chi^2_{0.05}(4-1) \quad$ qchisq$(0.95, 3) = 7.815$

Because observed $Q = 6.25 < 7.815$ we do not reject H_0.

Example 2: Flip 3 coins at same time and record the number of heads $X = 0, 1, 2, 3$ If $P(head) = 0.3$ and 3 coins are independently flipped, we should have $X \sim Binomial(3, 0.3)$. Suppose flip n=200 times, and observe $\underset{O_1}{Y_0 = 57}$, $\underset{O_2}{Y_1 = 95}$, $\underset{O_3}{Y_2 = 38}$, $\underset{O_4}{Y_3 = 10}$.

$H_0 : X \sim Binomial(3, 0.3)$ v.s. $H_1 : X$ follows other distributions. $\alpha = 0.05$.

$$
H_0 \left\{ \begin{array}{l}
P_1 = P_{10} = P(X=0) = \binom{3}{0} 0.3^0 0.7^3 = 0.343 \\
P_2 = P_{20} = P(X=1) = \binom{3}{1} 0.3^1 0.7^2 = 0.441 \\
P_3 = P_{30} = P(X=2) = \binom{3}{2} 0.3^2 0.7^1 = 0.189 \\
P_4 = P_{40} = P(X=3) = \binom{3}{3} 0.3^3 0.7^0 = 0.027
\end{array} \right.
$$

$E_1 = 200 \times P_{10} = 68.6$

$E_2 = 200 \times P_{20} = 88.2$

$E_3 = 200 \times P_{30} = 37.8$

$E_4 = 200 \times 0.027 = 5.4$

double check the calculation $\sum_i E_i = n$

$$Q = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(57-68.6)^2}{68.6} + \frac{(95-88.2)^2}{88.2} + \frac{(38-37.8)^2}{37.8} + \frac{(10-5.4)^2}{5.4}$$

$$= 6.4$$

Critical Region $Q \geq \chi^2_{0.05}(4-1) = 7.815$

Do not reject H_0

Sometimes you are interested in testing whether a data set fits a probability model with d parameters left unspecified.

For instance what if the probability of head was unspecificed in the previous example and you simply want to know whether the distribution has a form of binomial?

1. Estimate the d parameters e.g. using the maximum likelihood method.
2. Calculate the chi-square statistic Q using the obtained estimates.
3. Compare the chi-square statistic to a chi-square distribution with $k-1-d$ degrees of freedom.

d is the number of unknown parameters that we need to estimate in order to get p_{i0}'s

Example 3: Flip 3 coins at same time and record the number of heads $X = 0, 1, 2, 3$. Suppose flip n=200 times, and observe $\underset{O_1}{Y_0} = 57$, $\underset{O_2}{Y_1} = 95$, $\underset{O_3}{Y_2} = 38$, $\underset{O_4}{Y_3} = 10$.

$H_0 : X \sim Binomial(3, p)$ where p is unknown v.s. $H_1 : X$ follows other distributions. $\alpha = 0.05$.

To estimate p, we use sample proportion $\hat{p}$

$200 \times 3 = 600$ flips $\quad \hat{p} = \dfrac{0\times 57 + 1\times 95 + 2\times 38 + 3\times 10}{600} = \dfrac{201}{600} = 0.335$

$$p_{10} = P(X=0) = \binom{3}{0} 0.335^0 (1-0.335)^3$$

similarly we calculate p_{20}, p_{30}, p_{40}

$$E_i = 200 \times p_{i0} \qquad Q = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}$$

we estimated $\hat{p} = 0.335$

critical region $\quad Q \geq \chi^2_{0.05}(4-1-1)$

$$\chi^2_{0.05}(2)$$

Sometimes, we also collapse categories with small probabilities. The chi-square distribution relies on C.L.T. which needs relatively large sample size for each category, e.g. $E_i \geq 5$.

Example 4: We observe the number of small particles and conducted 100 experiments. Let X be the number of particles in each experiment and Y be the frequency of getting each set of X values.

	×1 ×2 ×2								10×2 11×1 12×1
X	(0,1,2)	3	4	5	6	7	8	9	10 and more
Y	5	13	19	16	15	9	12	7	4
	1	2	3	4	5	6	7	8	9

$\sum_{i=1}^{9} P_i = 1$

Instead of combining $X = 10, 11, 12$ we combine $X \geq 10$

Test H_0 $X \sim \text{Poisson}(\lambda)$ λ is unknown at $\alpha = 0.05$

Step 1 λ is unknown we use the point estimate $\hat{\lambda} = \bar{x} = 5.59$

MLE for Poisson

Step 2 Find E_i p.m.f. $P(X=x) = \dfrac{\lambda^x}{x!} e^{-\lambda}$ $x = 0, 1, 2 \cdots$

$$P_{10} = P(X = 0,1,2 \mid X \sim \text{Poisson}(5.59))$$
$$= P(X=0) + P(X=1) + P(X=2) = 0.0824$$

$$P_{20} = P(X=3) = \frac{5.59^3 e^{-5.59}}{3!} = 0.1082$$

......

$$P_{90} = P(X \geq 10) = 1 - \sum_{i=1}^{8} P_{i0}$$

E_i : 8.24 10.82 15.15 16.02 15.84 12.67 8.87 5.52 5.39

$n \times P_{i0}$ double check $\sum_{i=1}^{9} E_i = n = 100$

$k - 1 - d$ # of free parameter

$$Q = \sum_{i=1}^{9} \frac{(Q_i - E_i)^2}{E_i} = 5.7157$$

critical region $Q \geq \chi^2_{0.05}(9-1-1) = \chi^2_{0.05}(7)$
$= 14.07$

We do not reject H_0