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: $p_1 = p_{10}, p_2 = p_{20}, \cdots, p_K = p_{K0}$,

H_1: at least one equation does not hold.

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

Extend to k categories:

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

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

$Q \sim \chi^2$ 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.

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

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

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

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 $Y_0 = 57$, $Y_1 = 95$, $Y_2 = 38$, $Y_3 = 10$.

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

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.

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 $Y_0 = 57$, $Y_1 = 95$, $Y_2 = 38$, $Y_3 = 10$.

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

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.

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