Lecture 9: More Tests

Chi-square Goodness-of-fit Tests

The chi-square test is based on a statistic that was developed by Karl Pearson in 1900. Consider $Y_1 \sim b(n, p_1)$, $0 < p_1 < 1$. By the Central Limit Theorem,

$$Z = \frac{Y_1 - np_1}{\sqrt{np_1(1-p_1)}} \approx N(0,1) \text{ for large } n,$$

$$np_1 \geq 5, \; n(1-p_1) \geq 5.$$

So $Q_1 = Z^2 \approx \chi^2(1)$. Let $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$.

Then $Q_1 = \dfrac{(Y_1 - np_1)^2}{np_1(1-p_1)} = \dfrac{(Y_1 - np_1)^2}{np_1} + \dfrac{(Y_1 - np_1)^2}{n(1-p_1)}$

Since $(Y_1 - np_1)^2 = (n - Y_1 - n[1-p_1])^2 = (Y_2 - np_2)^2$,

we have

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2}$$

But $E(Y_1) = np_1$ and $E(Y_2) = np_2$

$$\Rightarrow Q_1 = \sum_{i=1}^{2} (Y_i - np_i)^2 / np_i \approx \chi^2(1).$$

Q_1 measures the "distance" between the observed values of Y_1 and Y_2 to the corresponding expected values.

To generalize, suppose an experiment has k mutually exclusive and exhaustive outcomes, say $A_1, A_2, \ldots, A_k$. Let $p_i = P(A_i)$, $\sum_1^k p_i = 1$.

The experiment is repeated n independent times. Let Y_i = the number of times the experiment results in A_i, $i = 1, 2, \ldots, k$. The joint distribution of $Y_1, Y_2, \ldots, Y_{k-1}$ is a generalization of the binomial distribution (called the multinomial distribution). We ~~write~~ write

$$f(y_1, y_2, \ldots, y_{k-1}) = P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_{k-1} = y_{k-1})$$

where $\sum_1^{k-1} y_i \le n$ and $y_1, y_2, \ldots, y_{k-1}$ are non-negative integers (note that $Y_k = n - \sum_1^{k-1} Y_i$).

Because of independence of the trials, the probability of each particular arrangement of y_1 A_1s, y_2 A_2s, y_k A_ks is $p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$. and the number of such arrangements is the multinomial coefficient

$$\binom{n}{y_1, y_2, \ldots, y_k} = \frac{n!}{y_1!\, y_2! \cdots y_k!}$$

Hence the p.m.f of $Y_1, Y_2, \ldots, Y_{k-1}$ is

$$f(y_1, y_2, \ldots, y_{k-1}) = \frac{n!}{y_1!\, y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$$

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i} \approx \chi^2(k-1)$$

To apply this statistic, consider the test for

$H_0: p_i = p_{i0}, \; i = 1, 2, \ldots, k$ vs $H_a: p_i \neq p_{i0}$

We tend to favour H_0 if y_i and np_{i0} are approximately equal; that is, if

$$q_{k-1} = \sum_{i=1}^{k} (y_i - np_{i0})^2 / np_{i0} \quad \text{is "small".}$$

Since $Q_{k-1} \approx \chi^2(k-1)$, then we shall reject H_0 if $q_{k-1} \geq \chi^2_{\alpha}(k-1)$, where α is the desired significance level of the test.

Example 1:

Let X = # heads when four coins are tossed at random.

If the coins are fair, then $X \sim b(4, \frac{1}{2})$.

Suppose the outcomes are as follows:

x	0	1	2	3	4	Total
Trials	7	18	40	31	4	100

Do these results support the assumption of $b(4, \frac{1}{2})$ as a reasonable model for X?

Solution

Let $A_1 = \{0\}$, $A_2 = \{1\}$, $A_3 = \{2\}$, $A_4 = \{3\}$ and $A_5 = \{4\}$.

If $p_{i0} = P(X \in A_i)$, then

$$p_{10} = p_{50} = \binom{4}{0}\left(\tfrac{1}{2}\right)^4 = \frac{1}{16} = 0.0625$$

$$p_{20} = p_{40} = \binom{4}{1}\left(\tfrac{1}{2}\right)^4 = 4/16 = 0.25$$

$$p_{30} = \binom{4}{2}\left(\tfrac{1}{2}\right)^4 = \frac{6}{16} = 0.375$$

The null hypothesis $H_0: p_i = p_{i0}$, $i = 1, 2, 3, 4, 5$

is rejected if $Q_4 > \chi^2_{0.05}(4) = 9.488$

when $\alpha = 0.05$

But $$q_4 = \frac{(7-6.25)^2}{6.25} + \frac{(18-25)^2}{25}$$

$$+ \frac{(40-37.5)^2}{37.5} + \frac{(31-25)^2}{25}$$

$$+ (4-6.25)^2/6.25$$

$$= 4.47 < 9.488$$

Thus, we fail to reject H_0 and conclude that $b(4, 1/2)$ is a reasonable probabilistic model for X.

Now suppose p_{i0} were unknown and

H_0: $X \sim b(n, p)$, $0 < p < 1$ i.e p_{i0} are functions of an unknown parameter p.

We would need an estimate of p to use in the calculation of $Q_4 = \sum_{i=1}^{5} \frac{(Y_i - np_{i0})^2}{np_{i0}}$, which is still approximately χ^2-distributed.

This estimator of p will be chosen to minimize Q_4 and is known as the minimum chi-square estimator of p, $\tilde{p}$.

Here $p_{i0} = P(X \in A_i) = \frac{4!}{(i-1)!(5-i)!} p^{i-1}(1-p)^{5-i}$

$i = 1, 2, 3, 4, 5$

If $\tilde{p}$ is used in Q_4, then

$Q_4 \approx \chi^2(3)$. In general, $Q_4 \approx \chi^2(4-d)$

if p_{i0} are functions of d unknown parameters.

q_{k-1} values are hence compared with $\chi^2(k-1-d)$ in testing H_0. Normally, MLEs are used, since minimum chi-square estimators are difficult to find.

Now, let W be a random variable of the continuous type and $W \sim F(w)$ = distribution function.

To test $H_0: F(w) = F_0(w)$ (known), we partion $[0,1]$ into k sets with points $b_0, b_1, b_2, \dots, b_k$, where $0 = b_0 < b_1 < b_2 < \dots < b_k = 1$.

Let $a_i = F_0^{-1}(b_i)$, $i = 1, 2, \dots, k-1$;

$$A_1 = (-\infty, a_1], \quad A_i = (a_{i-1}, a_i], \; i = 1, 2, \dots, k-1$$

$$A_k = (a_{k-1}, \infty); \text{ and } p_i = P(W \in A_i)$$
$$i = 1, 2, \dots, k.$$

Let Y_i = # times the observed value of W belongs to A_i, $i = 1, 2, \dots, k$, in n independent repetitions of the experiment

Then $Y_1, Y_2, \dots, Y_k$ have a multinomial distribution with parameters $n, p_1, \dots, p_{k-1}$.

Let $p_{i0} = P(W \in A_i)$ under $H_0 : W \sim F_0(w)$.

Then $H_0' : p_i = p_{i0} \quad i = 1, 2, \dots, k$. We reject H_0 if

$$q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}} \geq \chi^2_{\alpha}(k-1).$$

<u>Example 2</u>

Let X = time (in minutes) between calls to 911. We wish to test $H_0 : X \sim exp(\theta = 20)$. The table below provides the summary of $n = 105$ such calls.

Class	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
Frequency	41	22	11	10	9	5	2	3	2
Probability	0.36	0.23	0.147	0.094	0.0599	0.0382	0.0244	0.0155	0.027
Expected	38.05	24.26	15.47	9.86	6.29	4.011	2.562	1.628	2.87

$A_1 = [0, 9]$, $A_2 = (9, 18]$, $A_3 = (18, 27]$, $A_4 = (27, 36]$,

$A_5 = (36, 45]$, $A_6 = (45, 54]$, $A_7 = (54, 63]$,

$A_8 = (63, 72]$, $A_9 = (72, \infty)$.

$q_8 = 4.6861$, p-value $= 0.7905 \Rightarrow$ the exponential is an extremely good fit for the data.

Note that we assumed $\theta = 20$ but could have also run the test with $\hat{\theta} = \bar{x}$ (the MLE of θ) and achieve the same result.

In fitting data to the Normal distribution, $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown, we would estimate μ and σ^2 by $\bar{x}$ and S^2, then partition the sample space $\{w: -\infty < x < \infty\}$ into k mutually disjoint sets $A_1, A_2, \dots, A_k$ and use $\hat{p}_{i0} = \int_{A_i} \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{(x-\bar{x})^2}{2s^2}\right] dx$

$i = 1, 2, \dots, k$. Using observed frequencies $y_1, \dots, y_k$ and $\hat{p}_{10}, \hat{p}_{20}, \dots, \hat{p}_{k0}$, we obtain $Q_{k-1} \approx \chi^2(k-1-2)$

Exercise

Let $X = \#$ alpha particles emitted by barium-133 in one tenth of a second. An experimeter takes 50 observations of X with a Geiger counter in a fixed position and partitions the set of outcomes into sets

$A_1 = \{0,1,2,3\}$, $A_2 = \{4\}$, $A_3 = \{5\}$,
$A_4 = \{6\}$, $A_5 = \{7\}$ and $A_6 = \{8, 9, 10, \ldots\}$

The sample mean number of particles, $\bar{x} = 5.4$.
The table below provides a summary of the data:

outcome	A_1	A_2	A_3	A_4	A_5	A_6	Total
Frequency	13	9	6	5	7	10	
Prob ($\hat{p}_{io}$)							
Expected							

Test $H_0: X \sim \text{Poisson}(\lambda)$. at $\alpha = 0.05$