Chi-square godness of fit test, chapter 4.10

Grethe Hystad

November 27, 2012

- We will use the Chi-square godness of fit test to test appropriateness of different probabilistic models.
- First let Y be $b(n, p_1)$, where $0 < p_1 < 1$. By the Central limit theorem, we have that

$$Z = \frac{Y_1 - np_1}{\sqrt{np1(1-p_1)}}$$

is approximately $N(0,1)$ for large n, in particular when $np_1 \geq 5$ and $n(1-p_1) \geq 5$. Hence $Q_1 = Z^2$ is approximately $\chi^2(1)$.

Let $Y_2 = n - Y_1$ and $p_2 = 1 - p_1$, then we have

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

We can write

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

so we have

$$Q_1 = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2} = \sum_{i=1}^{2} \frac{(Y_i - np_i)^2}{np_i}.$$

- Y_1 is the number of "successes" and np_1 is the expected number of "successes".
- Y_2 is the number of "failures" and np_2 is the expected number of "failures".
- Q_1 measures the "closeness" of the observed numbers to the corresponding expected numbers.

Generalization

- Suppose an experiment have k mutually exclusive exhaustive outcomes, say A_1, A_2,..., A_k.
- Let $p_i = P(A_i)$ and thus $\sum_{i=1}^{k} p_i = 1$.
- Repeat the experiment n independent times.
- Let Y_i be the number of times the experiment results in A_i for $i = 1, ..., k$.
- The joint p.m.f. of Y_1, Y_2..., Y_{k-1} is $f(y_1, y_2, ..., y_{k-1}) = \frac{n!}{y_1! y_2! \cdots y_k!} p_1^{y_1} p_2^{y_2} \cdots p_k^{y_k}$ with $y_k = n - y_1 - y_2 - \cdots - y_{k-1}$ which is a generalization of the binomial distribution.

Generalization

Let $Y_k = n - Y_1 - Y_2 \cdots - Y_{k-1}$. Define

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

which is approximately $\chi^2(k-1)$.

We would like to test whether $p_i = p(A_i)$ is equal to a known number p_{i0} for $i = 1, 2, ..., k$. The null hypothesis is

$$H_0 : p_i = p_{i0} \quad \text{for} \quad i = 1, 2, ..., k.$$

The alternative hypothesis is H_1: $p_i \neq p_{i0}$ for at least one $i = 1, 2, ..., k$.
We keep H_0 if the observed number of times that A_i occurred and the number of times A_i was expected to occur if H_0 were true, np_{i0}, are approximately equal.
Let y_i be the number of times A_i occurred. Thus, we keep H_0 if

$$q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i}$$

is "small". We reject H_0 if

$$q_{k-1} \geq \chi^2_\alpha(k-1).$$

Example

Toss a fair coin 3 times. Let X be the number of heads observed. Repeat the experiment 100 times. Suppose we obtained

x	frequency
0	12
1	35
2	38
3	15

Is $b(3, 0.5)$ a reasonable model for the distribution of X?

Solution is given in class.

R-code

```
> chisq.test(c(12,35,38,15),p=c(0.125,0.375,0.375,0.125))

        Chi-squared test for given probabilities

data:  c(12, 35, 38, 15)
X-squared = 0.6933, df = 3, p-value = 0.8748
```

- Now let the random variable be continuous. Let $F(w)$ be the distribution of W.
- We will test $H_0 : F(w) = F_0(w)$, where $F_0(w)$ is some known distribution function of the continuous type.
- Split W into k sets.
- First split $[0, 1]$ into k sets with the points $0 = b_0 < b_1 \cdot \cdot \cdot < b_k = 1$.
- Let $a_i = F_0^{-1}(b_i)$ for $i = 1, 2, ..., k - 1$.
- Let $A_1 = (-\infty, a_1)$, $A_i = (a_{i-1}, a_i]$ for $i = 2, 3, ..., k - 1$ and $A_k = (a_{k-1}, \infty)$. Let $p_i = P(W \in A_i)$ for $i = 1, 2, ..., k$.
- Let Y_i denote the number of times the observed values of W belongs to A_i for $i = 1, 2, ..., k$ in n independent repeats of the experiment.

- Then Y_1, Y_2,..., Y_k have a multinomial distribution with parameters n, p_1, p_2, p_{k-1}.
- Let $p_{i0} = P(W \in A_i)$ when the distribution function of W is $F_0(w)$.
- We test $H_0 : p_i = p_{i0}$ for $i = 1, 2, ..., k$.
- Define

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_{i0})^2}{np_{i0}}.$$

- Reject H_0 if $q_{k-1} \geq \chi^2_\alpha(k-1)$.
- If we do not reject H_0 above, then we do not reject $H_0 : F(w) = F_0(w)$.
- If there are d unknown parameters that need to be estimated from the observations, we then compare q_{k-1} to $\chi^2_\alpha(k-1-d)$.

Example 4.10-4

Example

The following table lists 105 observations of X, the times between calls to 911 in a small city. Suppose we know the mean $\theta = 20$. Test the null hypothesis that the distribution of X is exponential with a mean of $\theta = 20$. The numbers are:
30, 17, 65, 8, 38, 35, 4, 19, 7, 14, 12, 4, 5, 4, 2, 7, 5, 12, 50, 33, 10, 15, 2, 10, 1, 5, 30, 41, 21, 31, 1, 18, 12, 5, 24, 7, 6, 31, 1, 3, 2, 22, 1, 30, 2, 1, 3, 12, 12, 9, 28, 6, 50, 63, 5, 17, 11, 23, 2, 46, 90, 13, 21, 55, 43, 5, 19, 47, 24, 4, 6, 27, 4, 6, 37, 16, 41, 68, 9, 5, 28, 42, 3, 42, 8, 52, 2, 11, 41, 4, 35, 21, 3, 17, 10, 16, 1, 68, 105, 45, 23, 5, 10, 12, 17.

Solution is given in class.