Contigency tables, chapter 4.11

Grethe Hystad

November 29, 2012

We will test

- whether two or more multinomial distributions are equal.
- test for independence of attributes of classifications.

- Suppose that each of two independent experiments have k mutually exclusive and exhaustive outcomes A_1, A_2,..., A_k. Let $p_{ij} = P(A_i)$ for $i = 1, 2, ..., k$ and $j = 1, 2, ..., h$.
- Thus, p_{11}, p_{21}, p_{k1} are the probabilities of the events in experiment 1 and
- p_{12}, p_{22}, p_{k2} are the probabilities of the events in experiment 2.
- Let experiment 1 and experiment 2 be repeated n_1 and n_2 independent times, respectively.
- Let Y_{11}, Y_{21}, Y_{k1} be the frequencies of A_1, A_2,..., A_k for the n_1 independent trials of experiment 1.
- Let Y_{12}, Y_{22},..., Y_{k2} be the frequencies of A_1, A_2,..., A_k for the n_2 independent trials of experiment 2.
- Then $\sum_{i=1}^{k} Y_{ij} = n_j$ for $j = 1, 2$.

- We know from chapter 4.10 that

$$\sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}}$$

  for $j = 1, 2$ is approximately $\chi^2(k-1)$.
- Since experiment 1 is independent of experiment 2,

$$\sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}}$$

  is approximately chi-square with $k - 1 + k - 1 = 2(k-1)$ degrees of freedom.
- We want to test $H_0 : p_{i1} = p_{i2}$ for $i = 1, 2, ..., k$ against $H_1 : p_{i1} \neq p_{i2}$ for at least one $i = 1, 2, ..., k$.

- Usually p_{ij} are unknown.
- Under H_0, we estimate the $k-1$ probabilities $p_{i1} = p_{i2}$ for $i = 1, 2, ..., k-1$ by

$$\frac{Y_{i1} + Y_{i2}}{n_1 + n_2} \quad i = 1, 2, ..., k-1.$$

- The estimator of $p_{k1} = p_{k2}$ is given by

$$\frac{Y_{k1} + Y_{k2}}{n_1 + n_2} = 1 - \frac{Y_{11} + Y_{12}}{n_1 + n_2} - ... - \frac{Y_{(k-1)1} + Y_{(k-1)2}}{n_1 + n_2}.$$

- We have that

$$Q = \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{\left[Y_{ij} - n_j \left(\frac{Y_{i1}+Y_{i2}}{n_1+n_2} \right) \right]^2}{\frac{n_j(Y_{i1}+Y_{i2})}{n_1+n_2}}$$

  has an approximately chi-square distribution with $2(k-1) - (k-1) = (k-1)$ degrees of freedom.
- $k-1$ is the number of estimated parameters.
- The critical region for testing H_0 is $q \geq \chi^2_\alpha(k-1)$.

Example

Suppose we want to test if online instruction in a math course gives different result than a traditional class. 50 students are selected at random from each of two groups. At the end of the semester each student is assigned a grade of A,B,C,D, or F. We have the following data:

	A	B	C	D	F	Total
online	10	13	16	9	2	50
traditional	8	11	15	10	6	50

Test if the two groups of instructions methods give about the same grades at the 5% significance level.

R-code

```
> w=matrix(c(10,8,13,11,16,15,9,10,2,6),nrow=2)
> chisq.test(w)

        Pearson's Chi-squared test

data:  w
X-squared = 2.4738, df = 4, p-value = 0.6493
```

Extension to testing equality of h independent multinomial distributions

- Let $p_{ij} = P(Ai)$ for $i = 1, 2, ..., k$ and $j = 1, 2, ..., h$. Repeat the j^{th} experiment n_j independent times. Let Y_{1j}, Y_{2j},..., Y_{kj} denote the frequency of the events A_1, A_2,..., A_k, respectively.
- $$\sum_{j=1}^{h}\sum_{i=1}^{k}\frac{(Y_{ij} - n_j p_{ij})^2}{n_j p_{ij}}$$

  has an approximately chi-square distribution with $h(k-1)$ degrees of freedom.

- Under H_0, we estimate $k-1$ probabilities using

$$\hat{p}_i = \frac{\sum_{j=1}^{h} Y_{ij}}{\sum_{j=1}^{h} n_j} \qquad i = 1, 2, ..., k-1.$$

- The estimate of $\hat{p}_k$ follows from $\hat{p}_k = 1 - \hat{p}_1 - ... - \hat{p}_{k-1}$.
- Thus,

$$\sum_{j=1}^{h}\sum_{i=1}^{k} \frac{(Y_{ij} - n_j\hat{p}_i)^2}{n_j\hat{p}_i}$$

has an approximate chi-square distribution with $h(k-1) - (k-1) = (h-1)(k-1)$ degrees of freedoms.

- Suppose a random experiment results in an outcome that can be classified by two different attributes.
- Suppose that the first attribute can terminate in one and only one of k mutually exclusive and exhaustive events, A_1, A_2,..., A_k.
- Suppose that the second attribute can terminate in one and only one of k mutually exclusive and exhaustive events, B_1, B_2,..., B_k.
- Define $p_{ij} = P(A_i \cap B_j)$ for $i = 1, 2, ..., k$ and $j = 1, 2, ..., h$. There are kh events $A_i \cap B_j$.
- Repeat the random experiment n independent times.
- Let Y_{ij} denote the frequency of the event $A_i \cap B_j$.

- The random variable

$$Q_{k-1} = \sum_{j=1}^{h} \sum_{i=1}^{k} \frac{(Y_{ij} - np_{ij})^2}{np_{ij}}$$

  is approximate $\chi^2(kh-1)$ provided n is large.
- We wish to test the hypothesis about independence of the A and B attributes.
- $H_0 : p(A_i \cap B_j) = p(A_i)p(B_j)$ for $i = 1, 2, ..., h$ and $j = 1, 2, ..., k$.
- Denote $p(A_i) = p_{i.}$ and $p(B_j) = p_{.j}$, where $p_{i.} = \sum_{j=1}^{h} p_{ij}$ and $p_{.j} = \sum_{i=1}^{k} p_{ij}$ and
- $1 = \sum_{j=1}^{h} \sum_{k=1}^{k} p_{ij}$.

- We will test
  $H_0 : p_{ij} = p_{i.}p_{.j}$ for $i = 1, 2, ..., k$ and $j = 1, 2, ..., h$.
- If $p_{i.}$, $p_{.j}$ are unknown, we estimate $\hat{p}_{i.} = \frac{y_{i.}}{n}$, where $y_{i.} = \sum_{j=1}^{h} y_{ij}$ is the observed frequency of A_i for $i = 1, 2, ...k$.
- We estimate $\hat{p}_{.j} = \frac{y_{.j}}{n}$, where $y_{.j} = \sum_{i=1}^{k} y_{ij}$ is the observed frequency of B_j for $j = 1, 2, ..., k$.
- We estimate $k - 1 + h - 1 = k + h - 2$ parameters.

- For $p_{ij} = p_{i.}p_{.j}$, we have

$$Q = \sum_{j=1}^{h}\sum_{i=1}^{k} \frac{[Y_{ij} - n(Y_{i.}/n)(Y_{.j}/n)]^2}{n(Y_{i./n})(Y_{.j}/n)}$$

  has a chi-square distribution with
  $df = kh - 1 - (k + h - 2) = (k - 1)(h - 1)$ d.f. provided H_0 is true.
- We reject H_0 if $Q \geq \chi^2_\alpha[(k - 1)(h - 1)]$.

Exercise 4.11-6

Example

A random sample of 50 women who were tested for cholesterol were classified according to age and cholesterol level and grouped in the following contigency table:

	Cholesterol level			
Age	< 180	180-210	> 210	Totals
< 50	5	11	9	25
≥ 50	4	3	18	25
Totals	9	14	27	50

Test the null hypothesis H_0 : Age and cholesterol level are independent attributes of classification. Use $\alpha = 0.01$.