VE401 Recitation Class Note11

Categorical Data

Chen Siyi
siyi.chen_chicy@sjtu.edu.cn

*Acceptance Sampling

Acceptance Sampling

A construction firm receives a shipment of N = 50 steel rods to be used in the construction of a bridge. The lot must be checked to ensure that the breaking strength of the rods meets specifications. The lot will be rejected if among the 200 rods more than 10% fail to meet specifications. We define the true proportion of defective for the 200 rods as Π. We test:

$$H_0 : \Pi \leq 0.1, \qquad H_1 : \Pi - 0.1 > \delta = 0.1$$

We test a sample of size n=20 and reject H_0 if more than c=3 rod fails to meet specifications.

1. What is α?
2. What is β?
3. *Can you draw the OC curve?
4. Keep α and n the same, if we want to make $\beta = 0.2$, what H_1 will you set?
5. To make $\alpha = 0.05$, assume n = 10, what value of c will you choose?
6. **To make $\alpha = 0.05$, $\beta = 0.3$, what value of n will you choose?

1 The Multinomial Distribution

The Multinomial Trial:
A multinomial trial with parameters p_1, ... , p_k is a trial that can result in exactly one of k possible outcomes. The probability that outcome i will occur on a given trial is p_i, for i = 1,...,k.

The Multinomial Random Variable:
A multinomial random variable now counts the number of times that outcome i occurs when a fixed number of n i.i.d. multinomial trials is performed. It therefore generalizes the binomial random variable.

The Multinomial Distribution:

$$(X_1, \ldots, X_k) : S \to \Omega = 0, 1, 2, \ldots, n^k$$

$$f_{X_1 X_2 \cdots X_k}(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Then $((X_1, \ldots, X_k), f_{X_1 X_2 \cdots X_k})$ follows a multinomial distribution with parameters n and p_1, ... , p_k.

Properties:

1. $\mathrm{E}[X_i] = \mathrm{n}p_i$
2. $\mathrm{Var}[X_i] = \mathrm{n}p_i(1\text{-}p_i)$
3. $\mathrm{Cov}[X_i, X_j] = \text{-}\mathrm{n}p_i p_j$

2 The Pearson Statistic

The Pearson Statistic:
For large n the Pearson statistic follows an approximate chi-squared distribution with k − 1 degrees of freedom.

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

Writing $O_i := X_i$ (observed frequencies) and $E_i := \mathrm{E}[X_i]$, (expected frequencies) it becomes

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Cochran's Rule:
We should require:

1. $\mathrm{E}[X_i] = np_i \geq 1$ for all $i = 1, \ldots, k$
2. $\mathrm{E}[X_i] = np_i \geq 5$ for 80% of all $i = 1, \ldots, k$

If not satisfied, combine categories!

3 Goodness-of-Fit Test

3.1 Pearson's Chi-squared Goodness-of-Fit Test

Goal:
Test whether a multinomial distribution have certain parameters $(p_{1_0}, \ldots, p_{k_0})$.

Method:
Let $(X_1, \ldots, X_k)$ be a sample of size n from a categorical random variable with parameters $(p_1, \ldots, p_k)$ **satisfying Cochran's Rule**. Let $(p_{1_0}, \ldots, p_{k_0})$ be a vector of null values. When n is large enough, we use the following statistic:

$$X^2_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_{i_0})^2}{np_{i0}}$$

which having the degree of freedom:

$$k - 1$$

To test:

$$H_0 : p_i = p_{i_0}, \quad i = 1, \ldots, k$$

We reject H_0 at significance level α if $X^2_{k-1} > \chi^2_{\alpha,k-1}$.
This simply means too large errors.

3.2 Goodness-of-Fit Test for a Distribution

Goal:

Test whether X follows certain discrete or continuous distribution with (unknown) m parameters.

*These estimated m parameters are parameters for X, while p_1, ... , p_k are parameters for the multinomial distribution $((X_1, \ldots, X_k), f_{X_1 X_2 \cdots X_k})$.

Method:

1. Find point estimates for the m unknown parameters
2. Using the estimated parameters, create categories. Make sure (X_1, ... , X_k) be a sample of size n from a categorical random variable with parameters (p_{1_0} , ... , p_{k_0}) **satisfying Cochran's Rule**.
3. When n is large enough, we use the following statistic:

$$X^2_{k-m-1} = \sum_{i=1}^{k} \frac{(X_i - np_{i_0})^2}{np_{i0}}$$

which having the degree of freedom:

$$k - m - 1$$

To test:

$$H_0 : \ X \ follows \ certain \ distribution.$$

We reject H_0 at significance level α if $X^2_{k-m-1} > \chi^2_{\alpha,k-m-1}$.

Goodness-of-Fit Test for a Distribution

During an experiment, test the number of α particles emitted by a kind of uranium During a fixed time period. Test for totally 100 times, and the results are in the table below. i is the number of α particles. f_i is the observed frequency. Is the test result consistent with reality?

i	0	1	2	3	4	5	6	7	8	9	10	11	⩾12
f_i	1	5	16	17	26	11	9	9	2	1	2	1	0
A_i	A_0	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}

4 Contingency Table Tests

4.1 Test for Independence

First we generate the contingency table

	column 1	column 2	...	column c	
row 1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
row 2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...
row r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	n

Then estimate probabilities according to:

$$\boxed{\widehat{p_{i\cdot}} = \frac{n_{i\cdot}}{n}, \quad \widehat{p_{\cdot j}} = \frac{n_{\cdot j}}{n}, \quad \widehat{p_{ij}} = \widehat{p_i} \cdot \widehat{p_{\cdot j}} = \frac{n_{i\cdot} n_{\cdot j}}{n^2}}$$

	column 1	column 2	...	column c	
row 1	$\widehat{p_{11}}$	$\widehat{p_{12}}$	...	$\widehat{p_{1c}}$	$\widehat{p_{1.}}$
row 2	$\widehat{p_{21}}$	$\widehat{p_{22}}$	...	$\widehat{p_{2c}}$	$\widehat{p_{2.}}$
...	...	...	...	...	...
row r	$\widehat{p_{r1}}$	$\widehat{p_{r2}}$	...	$\widehat{p_{rc}}$	$\widehat{p_{r.}}$
	$\widehat{p_{.1}}$	$\widehat{p_{.2}}$	...	$\widehat{p_{.c}}$	1

Then obtain E_{ij} according to:

$$\boxed{E_{ij} = n \cdot \widehat{p_{ij}} = \frac{n_{i\cdot} n_{\cdot j}}{n}}$$

	column 1	column 2	...	column c	
row 1	E_{11}	E_{12}	...	E_{1c}	$E_{1.}$
row 2	E_{21}	E_{22}	...	E_{2c}	$E_{2.}$
...	...	...	...	...	...
row r	E_{r1}	E_{r2}	...	E_{rc}	$E_{r.}$
	$E_{.1}$	$E_{.2}$	...	$E_{.c}$	n

Then we calculate the Pearson statistic:

$$\boxed{X^2_{(r-1)(c-1)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$$

which having the degree of freedom:

$$\boxed{(r-1)(c-1)}$$

because $k - 1 - m = rc - 1 - (r + c - 2) = (r - 1)(c - 1)$.
To test:

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$$

We reject H_0 at significance level α if $X^2_{(r-1)(c-1)} > \chi^2_{\alpha,(r-1)(c-1)}$.

Test for independence/homogeneity

A study is conducted to test for independence between air quality and air temperature. These data were obtained from records on 200 randomly selected days.
Is there an association between these variables?

	Air Quality		
Temperature	Poor	Fair	Good
Below average	1	3	24
Average	12	28	76
Above average	12	14	30

4.2 Test for Homogeneity

First we generate the contingency table

	column 1	column 2	...	column c	
row 1	n_{11}	n_{12}	...	n_{1c}	$n_{1.}$
row 2	n_{21}	n_{22}	...	n_{2c}	$n_{2.}$
...	...	...	...	...	...
row r	n_{r1}	n_{r2}	...	n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	...	$n_{.c}$	n

Then estimate probabilities according to:

$$\widehat{p_j} = \widehat{p_j} = \frac{n_{.j}}{n}$$

	column 1	column 2	...	column c	
row 1	$\widehat{p_1}$	$\widehat{p_2}$	...	$\widehat{p_c}$	$\widehat{p_{1.}} = 1$
row 2	$\widehat{p_1}$	$\widehat{p_2}$	...	$\widehat{p_c}$	$\widehat{p_{2.}} = 1$
...	...	...	...	...	...
row r	$\widehat{p_1}$	$\widehat{p_2}$	...	$\widehat{p_c}$	$\widehat{p_{r.}} = 1$

Then obtain E_{ij} according to:

$$\boxed{E_{ij} = n \cdot \widehat{p_{ij}} = n \cdot \widehat{p_j} = \frac{n_{i.} n_{.j}}{n}}$$

	column 1	column 2	...	column c	
row 1	E_{11}	E_{12}	...	E_{1c}	$E_{1.}$
row 2	E_{21}	E_{22}	...	E_{2c}	$E_{2.}$
...	...	...	...	...	...
row r	E_{r1}	E_{r2}	...	E_{rc}	$E_{r.}$
	$E_{.1}$	$E_{.2}$	...	$E_{.c}$	n

Then we calculate the Pearson statistic:

$$\boxed{X^2_{(r-1)(c-1)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$$

which having the degree of freedom:

$$\boxed{(r-1)(c-1)}$$

because $r(c-1) - (c-1) = (r-1)(c-1)$. $r(c-1)$ is the number of independent cells and $c-1$ is the number of independent parameters.

To test:

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj}, \quad j = 1, \ldots, c$$

We reject H_0 at significance level α if $X^2_{(r-1)(c-1)} > \chi^2_{\alpha,(r-1)(c-1)}$.