VE401 RC Week9

Wang Yangyang

UM-SJTU JI

2022 Spring

Outline

Outline

Categorical Random Variables

A random variable X that can take on the values $1, \ldots, k$ with respective probabilities $p_1 \ldots, p_k$ as above. A random sample of size n from X is collected and the results are expressed as a ***random vector***

$$(X_1, X_2, \ldots, X_k) \quad \text{with} \quad X_1 + X_2 + \cdots + X_k = n$$

The Multinomial Distribution: A random vector $((X_1, \ldots, X_k), f_{X_1 X_2 \cdots X_k})$ where

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

$p_1, \ldots, p_k \in (0,1), n \in \mathbb{N} \backslash \{0\}$ is said to have a multinomial distribution with parameters n and $p_1, \ldots, p_k$

- $\mathrm{E}[X_i] = np_i, \quad i = 1, \ldots, k$
- $\mathrm{Var}[X_i] = np_i(1 - p_i), i = 1, \ldots, k$
- $\mathrm{Cov}[X_i, X_j] = -np_i p_j, 1 \leqslant i < j \leqslant k$

The Pearson Statistics

Let $((X_1, \ldots, X_k), f_{X_1 X_2 \cdots X_k})$ be a multinomial random variable with parameters n and $p_1, \ldots, p_k$. For large n the *Pearson statistic*

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

follows an approximate chi-squared distribution with $k-1$ degrees of freedom (because we have $k-1$ independent cells).

Cochran's Rule: This tell us how large n needs to be for the chi-squared distribution to be a good approximation to the true distribution of the Pearson statistic when

$$\begin{aligned} \mathrm{E}[X_i] = np_i \geqslant 1, &\quad \text{for all } i = 1, \ldots, k \\ \mathrm{E}[X_i] = np_i \geqslant 5, &\quad \text{for 80\% of all } i = 1, \ldots, k \end{aligned}$$

Outline

Test for Multinomial Distribution

Let $(X_1, \ldots, X_k)$ be a sample of size n from a categorical random variable with parameters $(p_1, \ldots, p_k)$. We perform the chi-squared goodness-of-fit test.

Note: In this test, we directly make assumptions on parameters p_i ***without estimation based on samples***. This may happen when we already have some prior knowledge of the distribution (e.g. PRNG).

Procedures

1. Set
$$H_0 : p_i = p_{i_0}, \quad i = 1, \ldots, k$$
2. Calculate the expected values
$$E_i = np_{i0}$$

Then test whether the ***Cochran's rule*** is satisfied.

Test for Multinomial Distribution

Procedures

3. If satisfied, calculate the Pearson statistic.

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

which follows a chi-squared distribution with

degrees of freedom: *independent cells* $-m = k-1$

independent cells = k - 1, m=0

4. We reject H_0 at significance level α if $X_{k-1}^2 > \chi_{\alpha,k-1}^2$.

Goodness-of-Fit Test for Discrete Distribution

Now, we calculate the *estimates for parameters*, to make assumptions indirectly

Procedures

1. Suppose we guess that data *follow some distribution*, so we set the hypothesis as

   H_0 : A specific distribution with unknown parameter p_i

   e.g., H_0 : A Poisson distribution with parameter k.

2. Estimate parameters p_i from the sample based on your hypothesis.
   e.g., for *Poisson distribution*, estimate k by $\widehat{k} = \bar{X}$.
   Then we can calculate p_i. Suppose we have three categories $x = 0, x = 1, x \geqslant 2$, then

$$p_0 = P[X = 0] = \frac{e^{-\widehat{k}}\widehat{k}^0}{0!}, \quad p_1 = P[X = 1] = \frac{e^{-\widehat{k}}\widehat{k}^1}{1!},$$

$$p_2 = P[X \geqslant 2] = 1 - P[X = 0] - P[X = 1]$$

Goodness-of-Fit Test for Discrete Distribution

Procedures

3. Calculate the expected values $E_i = np_i$. Then make a table by yourself as below,

Category i	Exp. Frequency E_i	Obs. Frequency O_i
0	np_0	x_0
1	np_1	x_1
2	np_2	x_2
...	...	...

4. Test whether the *Cochran's Rule* is satisfied. If not satisfied, then go back to procedure (ii) and (iii) and change your number of categories.

Goodness-of-Fit Test for Discrete Distribution

Procedures

5. If (iv) satisfied, calculate the *Pearson statistic*.

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

now follows a chi-squared distribution with

$$\#\text{ independent cells } - m = k - 1 - m$$

degrees of freedom, where m is the *number of parameters that we estimate*. e.g., for the previous Poisson distribution test with 3 categories,

$$k = 3, m = 1$$

6. We reject H_0 at significance level α if X^2 exceeds the critical value.

Outline

Test for Independence

We define the marginal row and column sums

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$$

For a contingency table as below,

	column 1	column 2	column 3	
row 1	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
row 2	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
row 3	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Test for Independence

Procedures

1. If the hypothesis is that row and column categorizations are independent, then it should be the case that

$$H_0 : p_{ij} = p_i \cdot p_j$$

2. Estimates for the row and column probabilities are $\widehat{p_{i\cdot}} = \frac{n_{i\cdot}}{n}, \widehat{p_{\cdot j}} = \frac{n_{\cdot j}}{n}$, so if H_0 is assumed,

$$\widehat{p_{ij}} = \widehat{p_{i\cdot}} \cdot \widehat{p_{\cdot j}} = \frac{n_i \cdot n_{\cdot j}}{n^2}$$

3. Calculate the expected values

$$E_{ij} = n \cdot \widehat{p_{ij}} = \frac{n_i \cdot n \cdot j}{n}$$

Then test whether the *Cochran's rule* is satisfied.

Test for Independence

Procedures

4. If satisfied, calculate the Pearson statistic.

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

which follows a *chi-squared distribution* with degrees of freedom:

$$\#\text{independent cells } -m = rc-1-(r-1+c-1) = (r-1)(c-1)$$

- #independent cells $= rc-1$ because we have rc catagories.
- $m = r-1+c-1$ because we estimate $p_{i\cdot}$ with $i \leqslant r-1$ and $p_{\cdot j}$ with $j \leqslant c-1$.

5. We reject H_0 if the value of $X^2_{(r-1)(c-1)}$ exceeds the critical value.

Test for Comparing Proportions

Now the row totals are fixed, rewrite the table in terms of proportions:

	column 1	column 2	column 3	
row 1	p_{11}	p_{12}	p_{13}	$p_{1.} = 1$ (fixed)
row 2	p_{21}	p_{22}	p_{23}	$p_{2.} = 1$ (fixed)
row 3	p_{31}	p_{32}	p_{33}	$p_{3.} = 1$ (fixed)
row 4	p_{41}	p_{42}	p_{43}	$p_{4} = 1$ (fixed)

We want to compare proportions from each row, so

$$H_0 : \begin{cases} p_{11} = p_{21} = p_{31} = p_{41} \\ p_{12} = p_{22} = p_{32} = p_{42} \\ p_{13} = p_{23} = p_{33} = p_{43} \end{cases}$$

Test for Comparing Proportions

Procedure

1. Supposing that H_0 is true,

$$p_j := p_{1j} = p_{2j} = p_{3j} = p_{4j}$$

where p_j is the proportion of all objects following into the j th column. If H_0 is assumed, estimates for the column proportions are

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

2. Calculate the expected values

$$E_{ij} = n_i \cdot \widehat{p_{ij}} = n_i \cdot \widehat{p_j} = \frac{n_i \cdot n_{\cdot j}}{n}$$

Then test whether the *Cochran's rule* is satisfied.

Test for Comparing Proportions

Procedure

1. If satisfied, calculate the Pearson statistic.

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

which follows a ***chi-squared*** distribution with

$$\#\text{ independent cells } - m = r(c-1) - (c-1) = (r-1)(c-1)$$

degrees of freedom.

- #independent cells $= r(c-1)$ because only the number of objects in the first $c-1$ columns can be independently chosen, so we have a total of $r \cdot (c-1)$ independent cells.
- $m = (c-1)$ because we estimate p_j for $c-1$ times.

2. We reject H_0 if the value of $X^2_{(r-1)(c-1)}$ exceeds the critical value.

Outline

Simple Linear Regression Model

We assume that

$$Y \mid x = \beta_0 + \beta_1 x + E,$$

where $E[E] = 0$. We want to find estimators

$$\begin{aligned} B_0 := \widehat{\beta_0} &= \text{ estimator for } \beta_0, \quad b_0 = \text{ estimate for } \beta_0, \\ B_1 := \widehat{\beta_1} &= \text{ estimator for } \beta_1, \quad b_1 = \text{ estimate for } \beta_1. \end{aligned}$$

Assumptions

- For each value of x, the random variable follows a normal distribution with variance σ^2 and mean $\mu_{Y|x} = \beta_0 + \beta_1 x$.
- The random variables $Y \mid x_1$ and $Y \mid x_2$ are independent if $x_1 \neq x_2$.

Least Square Estimator

We have the error ***sum of squares***

$$\mathrm{SS_E} := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

To ***minimize*** it, we take

$$\frac{\partial \mathrm{SS_E}}{\partial b_0} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial \mathrm{SS_E}}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)\, x_i = 0$$

which gives

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}$$

Properties

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y}) y_i = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i = \sum_{i=1}^{n} (y_i - \bar{y}) x_i =$$

$$\sum_{i=1}^{n} x_i y_i - n\bar{x} \cdot \bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right).$$

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}, \qquad \mathrm{SS}_E = S_{yy} - b_1 S_{xy}.$$

Calculation Procedure

1. Find $\sum x_i, \sum y_i, \sum x_i^2, \sum y_i^2, \sum x_i y_i$ and calculate

$$S_{xx} = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2, \quad S_{yy} = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$
$$S_{xy} = \sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)$$

2. Obtain b_1 and b_0 by

$$b_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

3. Calculate other quantities as required, e.g.,

$$\mathrm{SS_T} = S_{yy}, \quad \mathrm{SS_E} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}, \quad R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

Outline

Distribution of Estimator for Variance

An unbiased estimator for variance σ^2 is given by

$$S^2 = \frac{\mathrm{SS_E}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_{Y|x_i}\right)^2$$

The statistic

$$\chi^2_{n-2} = \frac{(n-2)S^2}{\sigma^2} = \frac{\mathrm{SS_E}}{\sigma^2}$$

follows a ***chi-squared distribution*** with $n-2$ degrees of freedom.

Distribution of B_1 with Estimated Variance

The least squares estimator B_1 for β_1 follows a normal distribution with

$$\mathrm{E}[B_1] = \beta_1, \quad \mathrm{Var}[B_1] = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}.$$

The statistics

$$T_{n-2} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

follows *T-distributions* with $n-2$ degrees of freedom.
The $100(1-\alpha)\%$ *confidence interval* of β_1 is given by

$$B_1 \pm t_{\alpha/2,n-2} \frac{S}{\sqrt{S_{xx}}}.$$

Distribution of B_0 with Estimated Variance

The least squares estimator B_0 for β_0 follows a normal distribution with

$$\mathrm{E}[B_0] = \beta_0, \quad \mathrm{Var}[B_0] = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

The statistics

$$T_{n-2} = \frac{B_0 - \beta_0}{S\sqrt{\sum x_i^2}/\sqrt{nS_{xx}}}$$

follows *T-distributions* with $n-2$ degrees of freedom.
The $100(1-\alpha)\%$ *confidence interval* of β_0 is given by

$$B_0 \pm t_{\alpha/2,n-2} \frac{S\sqrt{\sum x_i^2}}{\sqrt{nS_{xx}}}.$$

Distribution of Estimated Mean

The estimated mean $\widehat{\mu}_{Y|x}$ follows a normal distribution with mean and variance

$$\mathrm{E}\left[\widehat{\mu}_{Y|x}\right] = \mu_{Y|x}, \quad \mathrm{Var}\left[\widehat{\mu}_{Y|x}\right] = \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\sigma^2.$$

Therefore, the statistic

$$T_{n-2} = \frac{\widehat{\mu}_{Y|x} - \mu_{Y|x}}{S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$$

follows a *T-distribution* with $n-2$ degrees of freedom. A $100(1-\alpha)\%$ *confidence interval* for $\mu_{Y|x}$ is given by

$$\widehat{\mu}_{Y_x} \pm t_{\alpha/2,n-2} S\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}.$$

Distribution and CI for Predictor

The statistic $Y \mid x - \widehat{Y \mid x}$ follows a normal distribution with mean and variance

$$\mathrm{E}[Y \mid x - \widehat{Y \mid x}] = 0, \quad \mathrm{Var}[Y \mid x - \widehat{Y \mid x}] = \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right) \sigma^2.$$

Therefore, the statistic

$$T_{n-2} = \frac{Y \mid x - \widehat{Y \mid x}}{S\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}}$$

follows a *T-distribution* with $n - 2$ degrees of freedom. A $100(1 - \alpha)\%$ ***prediction** interval* for $Y \mid x$ is given by

$$\widehat{Y \mid x} \pm t_{\alpha/2, n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

Outline

Quantities

1. *Total sum of squares*:

$$\mathrm{SS_T} = S_{yy} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2.$$

2. *Error sum of squared*:

$$\mathrm{SS_E} = \sum_{i=1}^{n} \left(Y_i - (B_0 + B_1 x_i)\right)^2 = S_{yy} - B_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

3. *Coefficient of determination*: the proportion of the total variation in Y that is explained by the linear model.

$$R^2 = \frac{\mathrm{SS_T} - \mathrm{SS_E}}{\mathrm{SS_T}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

Test for Significance with B_1

Let $(x_i, Y \mid x_i), i = 1, \ldots, n$ be a random sample from $Y \mid x$. We reject

$$H_0 : \beta_1 = 0$$

at significance level α if the test statistic

$$T_{n-2} = \frac{B_1}{S/\sqrt{S_{xx}}}$$

satisfies $|T_{n-2}| > t_{\alpha/2,n-2}$.

Test for Significance with R^2

Let $(x_i, Y \mid x_i)$, $i = 1, \ldots, n$ be a random sample from $Y \mid x$. We reject

$$H_0 : \beta_1 = 0$$

at significance level α if the test statistic

$$T_{n-2} = \frac{B_1}{S/\sqrt{S_{xx}}} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

satisfies $|T_{n-2}| > t_{\alpha/2,n-2}$.

Test for Correlation with R^2

Let (X, Y) follow a ***bivariate normal distribution*** with correlation coefficient $\rho \in (-1, 1)$. Let R be the estimator for ρ. Then we reject

$$H_0 : \rho = 0$$

at significance level α if the test statistic

$$T_{n-2} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

satisfies $|T_{n-2}| > t_{\alpha/2,n-2}$.

Testing for Lack of Fit

$\mathrm{SS_E}$ is the variance of Y explained by the model.

- Error sum of squares due to pure error.

$$\mathrm{SS_{E,pe}} := \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} Y_{ij}\right)^2.$$

  The statistic $\mathrm{SS_{E,pe}}/\sigma^2$ follows a ***chi-squared distribution*** with $\sum_{i=1}^{k}(n_i - 1) = n-k$ degrees of freedom.

- Error sum of squares due to lack of fit:

$$\mathrm{SS_{E,ff}} := \mathrm{SS_E} - \mathrm{SS_{E,pe}}.$$

  The statistic $\mathrm{SS_{E,If}}/\sigma^2$ follows a ***chi-squared*** distribution with $k-2$ degrees of freedom.

Testing for Lack of Fit

Test for lack of fit. Let $x_1, \ldots, x_k$ be regressors and $Y_{i1}, \ldots, Y_{in_i}$, $i = 1, \ldots, k$ the measured responses at each of the regressors. Let $\mathrm{SS}_{\mathrm{E,pe}}$ and $\mathrm{SS}_{\mathrm{E,\ lf}}$ be the ***pure error*** and ***lack-of-fit sums of squares*** for a linear regression model. Then we reject at significance level α.

H_0 : the linear regression model is appropriate if the test statistic

$$F_{k-2,n-k} = \frac{\mathrm{SS}_{\mathrm{E,f}}/(k-2)}{\mathrm{SS}_{\mathrm{E,pe}}/(n-k)}$$

satisfies $F_{k-2,n-k} > f_{\alpha,k-2,n-k}$.

Outline

It is suggested that you solve problems in assignments using Mathematica. It's the best way to prepare for Final Exam.
This notebook file, for your reference, is credited to previous TA Zhang Xingjian and Joy Dong.
It would be better to write your own notebook file.

End

Credit to Zhanpeng Zhou (TA of SP21)
Credit to Fan Zhang (TA of SU21)
Credit to Liying Han (TA of SP21)
Credit to Zhenghao Gu (TA of SP20)