STAT 231

December 2, 2016

Review video: Dec 8, 2016

OH: on Learn

Roadmap

- Goodness of fit tests.
- Test for independence of categorical variables.
- Test of equality of proportions
- Final few points of Statistical Inference
- Examples / Applications from real data.

TESTS FOR GOODNESS OF FIT.

Objective: To test whether the data set that we have follows a certain distribution

$X_1, \ldots \ldots X_n$

Sample $\{x_1, \ldots \ldots x_n\}$

$$H_0 : X_i \sim f(x_i ; \theta)$$

θ = vector of parameters

We use the Likelihood Ratio Test Statistic

$$\Delta(\theta_0) = 2 \sum Y_i \log \frac{Y_i}{E_i}$$

Y_i = Observed frequency in category i

E_i = Expected frequency in category i, under the ass^n that H_0 is true.

Result:

<u>Result</u> $\Lambda(t_0) \sim \chi^2_{n-k-1}$

$df = n-k-1$

n = # of categories

k = # of parameters of θ we had to estimate under H_0

Example

$X_1, \ldots X_n$ $\quad$ $\left.\begin{array}{l} n = 50. \\ \bar{x} = 200 \end{array}\right\}$

$$H_0 : X_i \sim \mathrm{Exp}(\theta).$$

$$i = 1, \ldots n.$$

Step 0 :

Divide the observations into different groups, and compute the observed frequency of each group.

	Y_i	e_i
$[0, 20)$	y_1	e_1
$[20, 50)$	y_2	e_2
$[50, 60)$	y_3	e_3
≥ 60	y_4.	e_4

Check

Your data set is large and $y_j \geq 5$ $\forall j$

Step 1: Find the MLE of θ under the null hypothesis.

$$\hat{\theta} = \bar{x} = 200$$

Step 2: Estimate the probability of each category

$$P(\text{category } 2) = \int_{20}^{50} \frac{1}{\theta} e^{-x/\theta} \, dx$$

$$= \int_{20}^{50} \frac{1}{200} e^{-x/200} \, dx$$

Step 3: Calculate the expected frequencies e_i

n = sample size.

$$e_i = n \times p_i$$

Step 4.

Compute $\lambda(\theta_0) = 2 \sum y_i \cdot \ln y_i / e_i$.

value of the test-statistic.

Step 5: Compute the p-value.

$$p\text{-value} = P(\Lambda \geq \lambda(\theta_0))$$

$$\Lambda \sim \chi^2_2$$

$n = 4$

$k = 1$

$X_1, \dots, X_n.$

σ

$H_0: X_i \sim G(\mu, \cancel{9})$

$(a, b]$	
$(b, c]$	
$(c, d]$	
$(d, e]$	
$(e, f]$	
$> f$	

$\Lambda \sim \chi^2_4$

$n = 6$

$k = 1$

Test of independence of attributes :

Objective : The data is divided into two categories. ~~Check~~ Test whether the two categories are independent.

Examples. : To test whether there is an association between CS major and having a Canadian home town

Contingency table.

	Canad	Non Canadian	
Can. CS	35 y_{11}	43 y_{12}	78
Non CS'	18 y_{21}	69 y_{22}	87
	53	112	165

Relative Risk

$$\text{DATA SET} = \frac{y_{11}/y_{11}+y_{12}}{y_{21}/y_{21}+y_{22}}$$

$= 2.17$

Test for no association

		B Can	B^c Non	
A	CS	θ_{11}	θ_{12}	α_1
A^c	non CS	θ_{21}	θ_{22}	α_2
		β_1	β_2	

θ_{ij} = P(category i & j)

α_1 = P(CS major)

β_1 = P(Canadian home town)

No association implies

$$\underset{P(A \text{ and } B)}{\theta_{ij}} = \underset{P(A)}{\alpha_i} \cdot \underset{P(B)}{\beta_j} \quad \forall\ i, j$$

	CA	Non Can.		
CS	35	43	78	α_1
Non CS	18	69		$1-\alpha_1$
	53		165	
	β_1	$1-\beta_1$		

$H_0 \quad \theta_{ij} = \alpha_i \beta_j . \quad \left.\begin{matrix} i = 1, 2 \\ j = 1, 2 \end{matrix}\right\}$

We have to estimate

$\underbrace{\alpha_1 \quad \cancel{\alpha} \quad \beta_1}$

$\checkmark \hat{\alpha}_1 = ? \quad 78/165$

$\hat{\theta}_{11} = \frac{78}{165} \cdot \frac{53}{165}$

$\checkmark \hat{\beta}_1 = 53/165$

	Ca	No.
CS	$\hat{\theta}_{11}$	$\hat{\theta}_{12}$
NonC	$\hat{\theta}_{21}$	$\hat{\theta}_{22}$

Construct the Expected Frequency table

	Ca.	Non Cana.
CS	e_{11}	e_{12}
NonCS	e_{21}	e_{22}

$$e_{ij} = n \times \hat{\theta}_{ij}$$

Likelihood ratio

$$\Delta = 2 \sum_i \sum_j y_{ij} \log \frac{y_{ij}}{e_{ij}}$$

$df = ?$

$n = 4$ $\quad k = 2$

$$\boxed{\Delta \sim \chi^2_1}$$

o_{11} e_{11}

	C	NC	
CS	35 $\frac{78 \times 53}{165}$	43	78
Rural	18 $\frac{87 \times 53}{165}$	69	87
	53	112	165

$$e_{ij} = \frac{r_i \times c_j}{n}$$

Column Total.

Row Total

	Ca	Non.	
CS	θ_{11}	θ_{12}	α_1
ACTSC	θ_{21}	θ_{22}	α_2
Neither	θ_{31}	θ_{32}	α_3
	β_1	β_2.	1

$df = 6 - 1$

$R = 9$

$df = 6 - 3 - 1$
$= 2$

$$H_0 : \theta_{ij} = \alpha_i \beta_j \quad \forall i, j$$

$$e_{11} = \frac{r_1 \times c_1}{n}$$

r_1 = # of unit in Row 1

c_1 = # of ann. in Colm 1

$$df = (a-1)(b-1)$$

a = # of rows

b = # of columns..

Test for equality of proportions

	Male	Female
Smo		
Non		

$\equiv$ testing for independence