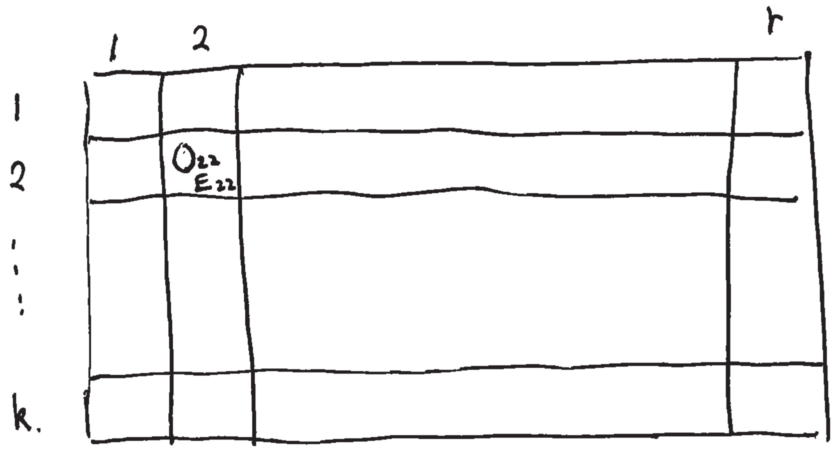
STAT 417 lecture Note 23

In previous lecture, we showed how to use χ^2 goodness fit test to test models. Today, we will talk about how to ~~use~~ test relationships among variables. We first look at a general form of

χ^2 - test:

Consider the following $k \times r$ cells. $(k, r \geq 2)$



n elements are put into these cells. each element goes to one cell.

Let O_{ij} = observed frequency of the cell C_{ij}.

E_{ij} = expected frequency of cell C_{ij}.

$i = 1, 2 \ldots k$.
$j = 1, 2 \ldots r$.

Then the χ^2 - test statistics is

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2((k-1)(r-1))$$

Theorem. $\chi^2 \sim \chi^2((k-1)(r-1))$. under various null hypothesis.

Next. we show how to use χ^2 - test to check relationships among variables.

<u>Eg 1</u>. (Independence checking). The most important relationship between variables is independence. This summer we have $n=10$ graduate students. from whom we get samples $(X_1, Y_1), \dots, (X_{10}, Y_{10})$.

where $X_i = \begin{cases} 1, & \text{if the } i\text{th student has taken STAT 417.} \\ 2. & \text{if o/w.} \end{cases}$

$Y_i = \begin{cases} 1 , & \text{doctor} \\ 2 & \text{lawyer} \\ 3 & \text{scientist} \\ 4 & \text{unemployed.} \end{cases}$

The 10 samples are below.

$(1,1)$, $(2,4)$. $(1,2)$, $(1,3)$, $(1,2)$, $(2,4)$. $(2,4)$, $(1,3)$,

$(2,4)$, $(1,3)$.

Summarizing the samples in a 2×4 cell.

	$y=1$	$y=2$	$y=3$	$y=4$	
$x=1$	1 O_{11}	2 O_{12}	3 O_{13}	0 O_{14}	6
$x=2$	0 O_{21}	0 O_{22}	0 O_{23}	4 O_{24}	4
	1	2	3	4	

Question: Does "taking STAT 417" affect your job hunting?

To test this, we just have to check whether X and Y are independent. The observed frequency in each cell is easy to collect.

Next we should find the expected frequency under independence assumption.

that is, the E_{ij}.

If X and Y are independence. then

$p_{11} = p(X=1, Y=1) = p(X=1)p(Y=1) = (0.6)(0.1) = 0.06$

$p_{12} = p(X=1, Y=2) = p(X=1)\ p(Y=2) = (0.6)(0.2) = 0.12$

$p_{13} = p(X=1, Y=3) = p(X=1)\ p(Y=3) = (0.6)(0.3) = 0.18$

$p_{14} = p(X=1, Y=4) = p(X=1)\ p(Y=4) = (0.6)(0.4) = 0.24.$

$p_{21} = p(X=2, Y=1) = p(X=2)\ p(Y=1) = (0.4)(0.1) = 0.04.$

$p_{22} = p(X=2, Y=2) = p(X=2)\ p(Y=2) = (0.4)(0.2) = 0.08$

$p_{23} = p(X=2, Y=3) = p(X=2)\ p(Y=3) = (0.4)(0.3) = 0.12$

$p_{24} = p(X=2, Y=4) = p(X=2)\ p(Y=4) = (0.4)(0.4) = 0.16$

So $E_{11} = n\,p_{11} = (10)(0.06) = 0.6.$

$E_{12} = n\,p_{12} = (10)(0.12) = 1.2.$

$E_{13} = n\,p_{13} = (10)(0.18) = 1.8.$

$E_{14} = n\,p_{14} = (10)(0.24) = 2.4.$

$E_{21} = n\,p_{21} = (10)(0.04) = 0.4$

$E_{22} = n\,p_{22} = (10)(0.08) = 0.8$

$E_{23} = n\,p_{23} = (10)(0.12) = 1.2.$

$E_{24} = n\,p_{24} = (10)(0.16) = 1.6.$

The χ^2-test is

$$\chi^2 = \frac{(1-0.6)^2}{0.6} + \frac{(2-1.2)^2}{1.2} + \frac{(3-1.8)^2}{1.8} + \frac{(0-2.4)^2}{2.4}$$

$$+ \frac{(0-0.4)^2}{0.4} + \frac{(0-0.8)^2}{0.8} + \frac{(0-1.2)^2}{1.2} + \frac{(4-1.6)^2}{1.6}$$

$$= 10 > \chi^2_{0.95}(3) = 7.81.$$

So. STAT 417 has significant impact on your future ~~car~~ job hunting.

<u>EX1</u>. Suppose ~~[illegible]~~ $Y \in \{0, 1\}$ is binary variable.

$X \in \{1, 2, 3, 4\}$ is categorical variable.

$n=100$ samples were collected. and is summarized in the following

2×4 table.

12	10	16	14	52
13	15	9	11	48
25	25	25	25.	

Test if X and Y are independent.

Sol: $E_{11} = E_{12} = E_{13} = E_{14} = (0.25)(0.52)(100) = 13.$

$E_{21} = E_{22} = E_{23} = E_{24} = (0.25)(0.48)(100) = 12.$

so $$\chi^2 = \frac{(12-13)^2}{13} + \frac{(10-13)^2}{13} + \frac{(16-13)^2}{13} + \frac{(14-13)^2}{13}$$

$$+ \frac{(13-12)^2}{12} + \frac{(15-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(11-12)^2}{12} = 3.21. < \chi^2_{0.95}(3) = 7.81.$$

so there is evidence that x and y are independent.