STAT 135 Lecture 14

Henry Liev

24 July 2025

Remark 0.1 (Test of independence)

"Nice application" of goodness of fit

Example 0.2 (Two categorical variables)

Survey about taste in music and sports

Sport	Pop	Hiphop	Country	Metal	Jazz
Basketball	40	50			
Football	20				
Soccer					
Baseball					

Is there an association?
Or is there no association? (independence)
H_0: Variables are independent
H_1: Not independent

Remark 0.3

Table with I rows and J columns, think of this as multinomial with IJ possible outcomes
Probability of being in cell $ij = \pi_{ij}$
Let P_i denote the probability of being in role i and q_j be the probability of being in column j
$q_j = \sum \pi_i$
$H_0 : \pi_{ij} = p_i q_j$
H_1 : Not independent

Method 0.4

Data n_{ij} = count in cell ij
Under H_0 we can think of
$i - 1$ row probabilities $p_1, p_2, \ldots, p_I$
$j - 1$ column probabilities $q_1, q_2, \ldots, q_J$
Think of the unknown p's and q's as parameters
$\pi_{ij}(p_i, q_j) = p_i q_j$

Method 0.5

Proceed "as usual"

1. Estimate the p_i and q_j by MLE
   $\hat{p}_i = \frac{n_{i.}}{n}$ $\qquad \hat{q}_j = \frac{n_{.j}}{n}$
2. Under H_0 : Expected count in cell $ij = n\hat{p}_i\hat{q}_j$
3. Chi squared statistic

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{O_{ij}-E_{ij}}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(n_{ij}-\frac{n_i n_j}{n}\right)^2}{\frac{n_i n_j}{n}}$$

$T \sim \chi^2_d \to T \sim \chi^2_{n-1-r} \sim \chi^2_{IJ-1-(I-1)-(J-1)} = \chi^2_{(I-1)(J-1)}$

	1	$\cdots$	J
1	n_{ij}		
$\vdots$			
I			

I rows

J columns

Let $n_{i.} = \sum_{\forall j} n_{ij}$ $\qquad n_{.j} = \sum_{\forall i} n_{ij}$

Example 0.6 (Chi Squared Test of Homogeneity)

4 7-sided dice (not fair)

1	2	3	4	5	6	7
p_1	p_2	p_3	p_4	p_5	p_6	p_7

Side	Die 1	Die 2	Die 3	Die 4
1	117			
2	25			
3	$\vdots$			
$\vdots$	$\vdots$			
7	100	250	70	90

Question: Are all the dice "the same?"

$I = \#$ sides, $J = \#$ dice

$H_0 : \pi_{i1} = \pi_{i2} = \cdots = \pi_{ij}$ for each $i, i = 1, \cdots, I$

Solution. J multinomial distributions with I possible outcomes

Under H_0 denote the common π_{ij} by π_i

$\pi_1 = P(X = 1) = \frac{n_{1.}}{n}$

$\pi_2 \vdots$

$\pi_I = \frac{n_{I.}}{n}$

$n_{i.} =$ count for side i, $n_{.j} =$ count for die j, $n_{ij} =$ count for side i on die j

Pick a j

$$T_j = \sum_{i=1}^{n}\frac{\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \qquad T_j \sim \chi^2_{I-1}$$

Sum up over the columns (dice)

$$T = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$$

$$T \sim \chi^2_{J(I-1)-(I-1)} = \chi^2_{(I-1)(J-1)}$$

■

Remark 0.7 (Homogeneity or Independence)

- Questions are very similar
- Is outcome independent of die label?
- Only real difference is the sampling model
- Analysis turns out the same either way

Method 0.8 (Non-parametric two sample tests)

Don't want to make any assumptions and don't have large samples
Suppose we have
$X_1, \ldots, X_n \overset{iid}{\sim} F$
$Y_1, \ldots, Y_m \overset{iid}{\sim} G$
Consider:
$H_0 : F = G$

1. Group all $m + n$ observations together
2. Sort them by increasing size
3. Just look at ranks
4. Look at sum of ranks of control
5. Reject H_0 if sum of ranks too extreme
6. Look at all $\binom{m+n}{n}$ ways to assign ranks, all equally likely under H_0
7. Get p-value based on where observed falls