Ch. 12 Analysis of Categorical Data and Goodness of Fit Tests

Goodness of Fit

Given a sample with k categories, tests how well the sample fits the hypothesized distribution.
Assumptions:
–Random sample and all expected counts are greater than 5.

1. $H_0 : p_i = p_{i0}$ for all i=1,2,...,k
2. $H_a : p_i \neq p_{i0}$ for at least one i=1,2,...,k
   Note: the alternative is always two-sided
3. Test statistic:
   $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
4. Find p-value
   pvalue = $P(\chi^2 > \chi^2_{obs})$, with df = k – 1
   pvalue is always the area in the upper tail region

Ex. Rolling a die 36 times, test the null hypothesis that the a die is a fair die given the observation:
1: 4 times; 2: 7 times; 3: 5 times; 4: 6 times; 5: 8 times; 6: 6 times

There are 6 categories,
$H_0 : p_i = 1/6$ for $i = 1, 2, ..., 6$
$H_a : p_i \neq 1/6$ for at least 1 $i = 1, 2, ..., 6$

If this is a fair die, we would expect 6 throws for each face.

$$\chi^2_{obs} = \frac{(4-6)^2}{6} + \frac{(7-6)^2}{6} + \frac{(5-6)^2}{6} + \frac{(6-6)^2}{6} + \frac{(8-6)^2}{6} + \frac{(6-6)^2}{6} = 1.67$$

The df = 6 – 1 = 5, pvalue = $P(\chi^2 > 1.67)$, using Table 8, pvalue > .1
Since the pvalue is less than α, we fail to reject the null hypothesis and conclude that the die is fair.

Chi-square test of homogeneity in a contingency table:

Year	CALS	Engineering	ILR	AS	Total
Freshman	30	40	20	40	130
Sophomore	20	50	30	40	140
Junior	30	20	35	55	140
Senior	35	25	15	30	105
Total	115	135	100	165	515

Test at 5% significance level that the distribution of student's affiliated college is evenly distributed among class year.

$H_0 : p_{1j} = p_{2j} = ... = p_{rj}$, for all j

H_a : at least one p_{ij} differs

First generate the expected values table using the formula: $E_{ij} = \frac{\text{RowTotal} * \text{ColumnTotal}}{\text{GrandTotal}}$

Year	CALS	Engineering	ILR	AS	Total
Freshman	29	34	25	42	130
Sophomore	31	37	27	45	140
Junior	31	37	27	45	140
Senior	23	28	20	34	105
Total	114	136	99	166	515

Due to rounding, the row and column totals may be slightly off. Note that all expected values are greater than 5.

Next, calculate χ^2_{obs}.

$\chi^2_{\text{obs}} = \frac{(30-29)^2}{29} + \frac{(40-34)^2}{34} + \frac{(20-25)^2}{25} + ...$, repeat for each cell of the table, so in total, you should have the summation of 16 fractions.

Next, find p-value. The df $= (\text{r}-1)(\text{k}-1) = (4-1)(4-1) = 9$, where r is the number of rows and k is the number of columns.

Lastly, compare p-value to alpha.

The chi-square test for independence:

- H_0 : The two categorical variables are independent.
- H_a : The two categorical variables are not independent.
- Test statistic: χ^2_{obs} (same formula)
- pvalue = $\text{P}(\chi^2_{\text{obs}} > \chi^2_{\text{obs}})$, and df = (number of rows - 1)*(number of columns - 1)
- Compare to α.