Ch. 26 - Comparing Counts

Notation: k = # of categories of a qualitative variable

p_i = true proportion of category i; $\quad i = 1,\ldots, k \quad$ (Note: $\sum_{i=1}^{k} p_i = 1$)

A random sample of size n will provide sample statistics of "observed counts". These values can compare against "expected counts" of np_i for each category. Consequently, an H_0 can collectively test the validity of each p_i. How?

Def'n: The "goodness-of-fit" test uses the chi-square statistic, χ^2, is computed by

$$\chi^2 = \sum_{cells} \frac{(Obs - Exp)^2}{Exp}$$

where *Obs* = "observed count", *Exp* = "expected count", and you sum over all categories. Sizeable differences between *Obs* and *Exp* of specific categories lead to large values of χ^2 and subsequent rejection of H_0. For formal rejection/non-rejection, we need a formal test.

Aside: The chi-squared distribution has the following properties:

- like the *t*-distribution, it has only one parameter, *df*, that can take on any positive integer value.
- skewed to the right for small *df* but becomes more symmetric as *df* increases.
- curve where all areas correspond to nonnegative values.
- values denoted by χ^2

When H_0 is correct and n sufficiently large, χ^2 approx. follows a χ^2-dist'n with $df = k - 1$. Using this dist'n, the corresponding *P*-value is the area to the right of χ^2 under the χ^2_{k-1} curve (all curves found in Appendix Table X). For test validity, the following must hold:

1) Observed cell counts are based on a random sample.
2) The sample size is large (every expected count ≥ 5).

Ex26.1) Table 26X0 - Number of Films in 2012 by Film Rating

Film Rating	Frequency (*Obs*)	Expected count (*Exp*)
G	15	$np_G = 443(0.25) = 110.75$
PG	62	110.75
PG-13	145	110.75
R	221	110.75

Are film ratings evenly distributed among all the movies made in 2012? Use $\alpha = 0.05$.

Assumptions: Entire population of American films, not random sample. We will assume it, but cautiously. Positively, all expected counts are greater than 5, so the "goodness-of-fit" test is possible.

H_0: $p_G = 0.25$, $p_{PG} = 0.25$, $p_{PG\text{-}13} = 0.25$, $p_R = 0.25$

H_A: at least one p_i is not as claimed

$$\chi^2 = \frac{(15-110.75)^2}{110.75} + \frac{(62-110.75)^2}{110.75} + \frac{(145-110.75)^2}{110.75} + \frac{(221-110.75)^2}{110.75}$$
$$= 82.782 + 21.459 + 10.592 + 109.752$$
$$= 224.585$$

At $\chi^2_{k-1} = \chi^2_3$, 224.585 is higher than the largest value of 12.838, which has a *P*-value of 0.005. Thus, the *P*-value range is (0, 0.005). With this range and the given $\alpha = 0.05$, reject H_0. Conclusively, there is enough evidence that the film ratings are not evenly distributed.

Testing for Homogeneity

Def'n: A two-way frequency table (or a *contingency table*) summarizes categorical data. Each cell in the table is a particular combination of categorical values.

Marginal totals occur by extending the table to include the sums of each row and column. In addition, the grand total occurs.

Table 26X1 – 2-way table of responses

	Like Hockey (A)	Indifferent (B)	Dislike Hockey (C)	**Row Marginal Total**
Male (M)	15	11	6	**32**
Female (F)	22	13	6	**41**
Column Marginal Total	**37**	**24**	**12**	**73**

The test for homogeneity determines if the category proportions are the same for all the populations. The expected cell counts under homogeneity are:

$$\frac{\text{(row marginal total)(column marginal total)}}{\text{grand total}}$$

Thus, the above table could now look like:

Table 26X2 – Expanded 2-way table of responses

	Like Hockey (A)	Indifferent (B)	Dislike Hockey (C)	**Row Marginal Total**
Male (M)	15 (16.2)	11 (10.5)	6 (5.3)	**32**
Female (F)	22 (20.8)	13 (13.5)	6 (6.7)	**41**
Column Marginal Total	**37**	**24**	**12**	**73**

*expected values in parentheses

Accordingly, we can calculate a value for χ^2 for the entire table under an H_0 that assumes homogeneity. When H_0 is correct, χ^2 approximately follows a χ^2-distribution such that $df = (R - 1)(C - 1)$. For test validity, the following must hold:

1) The data consist of independently chosen random samples.
2) The sample size is large (every expected count ≥ 5). If not, combine rows or columns, if appropriate, to achieve satisfactory expected counts.

Ex26.2) Is there homogeneity of the proportions of "hockey appreciation" between males and females? Use $\alpha = 0.05$.

H_0: homogeneity of proportions H_A: some absence of homogeneity

"Independent random samples" not stated, but knowing how data was collected, the samples probably are. Using Table 26X2, all expected counts are greater than 5, so the test is possible.

$$\chi^2 = \frac{(15-16.2)^2}{16.2} + ... + \frac{(6-6.7)^2}{6.7} = 0.0916 + \ldots + 0.0812 = 0.387$$

At χ^2_2, $0.387 < 4.605$ (P-value of 0.100); hence, the P-value range is (0.1, 1). With this range and the given $\alpha = 0.05$, do not reject H_0. Thus, there may be homogeneity of the proportions.

Testing for Independence
Def'n: The test for independence determines whether an association exists between two categorical variables. Comparing to the homogeneity test, the main change is in H_0 (and H_A), which now states that the 2 variables are independent. Also, the 1st assumption is now that the observed counts are from a single random sample. The other assumption and calculations remain the same.

NOTE that **NOT** rejecting H_0 is the preferred choice here.

Ex26.3) Are the two variables (Gender and Hockey Appreciation) dependent?

H_0: G & HA are independent H_A: G & HA are dependent

Using Table 26X2 and calculations in Ex26.2), the entire testing process reveals weak evidence against H_0. Thus, the two variables may be independent.

Ch. 26 - Summary
- tests for homogeneity are used when the subjects in each of 2 or more independent samples are classified according to a single categorical variable.
- tests for independence are used when the subjects in a *single* sample are classified according to two categorical variables.
- in our case, a test for independence seems more appropriate.