Ch. 23 – 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).

Ex23.1) Are film ratings evenly distributed for U.S. movies made in 2016? Use $\alpha = 0.05$.

Table 23X0 – Number of Films in 2016 by Film Rating

Film Rating	Frequency (*Obs*)	Expected count (*Exp*)
G		
PG		
PG-13		
R		

Assumptions:

H_0:

H_A:

$\chi^2 =$

$df =$

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 23X1 – 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:

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

Thus, the table with expected values would look like:

Table 23X2 – 2-way table of expected counts

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

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.

Ex23.2) Is there homogeneity of the proportions of "hockey appreciation" between males and females? Use $\alpha = 0.05$.

Assumptions:

H_0: H_A:

$\chi^2 =$

$df =$

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.

Ex23.3) Are the two variables (Gender and Hockey Appreciation) dependent?

H_0: H_A:

Ch. 23 – 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.