Stat 401, section 14.3 Goodness of Fit – Homogeneity, Independence

notes by Tim Pilachowski

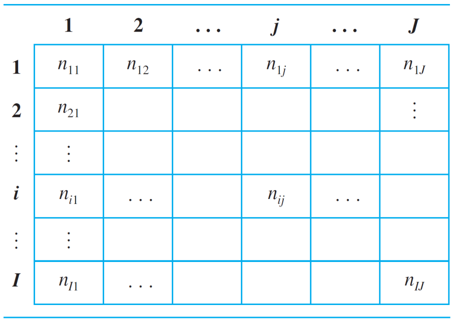
In Lecture 14.1 we considered a multinomial experiment with *k* possible outcomes and null hypothesis which specified the value of each p_i: $H_0: p_1 = p_{10},\ \ p_2 = p_{20},\ \ldots\ ,\ \ p_k = p_{k0}$. The test statistic was

$$\sum \frac{(\text{observed} - \text{expectd})^2}{\text{expectd}} = \sum \frac{(n_i - n\,p_{i0})^2}{n\,p_{i0}}$$, a chi-squared (χ^2) distribution with $\nu = k - 1$ degrees of freedom.

In Lecture 14.2 we looked at determining whether a set of data indicates that the population has an underlying normal distribution. Our process involved using maximum likelihood estimators based on the full sample $X_1, X_2, \ldots, X_n$, and letting χ^2 denote the statistic based on these estimators. We had two critical values, each with its own degrees of freedom, and the decision criteria was reject H_0 if $\chi^2 \ge \chi^2{}_{\alpha,k-1}$, fail to reject H_0 if $\chi^2 \le \chi^2{}_{\alpha,k-1-m}$, and withhold judgement if $\chi^2{}_{\alpha,k-1-m} \le \chi^2 \le \chi^2{}_{\alpha,k-1}$.

In section 14.3, we'll use a chi-squared (χ^2) test statistic to determine homogeneity and independence.

In this analysis, the data will consist of counts or frequencies (just as in sections 14.1 and 14.2), but the data will be displayed in a **two-way contingency table** which has *I* rows $(I \ge 2)$ and *J* columns, thus *IJ* cells.



	1	**2**	. . .	***j***	. . .	***J***
1	n_{11}	n_{12}	. . .	n_{1j}	. . .	n_{1J}
2	n_{21}					⋮
⋮	⋮					
i	n_{i1}	. . .		n_{ij}	. . .	
⋮	⋮					
I	n_{I1}	. . .				n_{IJ}

There are two commonly encountered situations in which such data arises:

1. There are *I* populations of interest (each corresponding to a different row of the table, and each population is divided into the same *J* categories. A sample is taken from the *i*th population (*i* = 1, 2, … , *I*) and the counts are entered in the cells in the *i*th row of the table. See Example A below.

In situations of type 1, we want to investigate whether the proportions in the different categories are the same for all populations. The null hypothesis states that the populations are *homogeneous* with respect to these categories.

$$H_0 : p_{1j} = p_{2j} = \ \ldots\ = p_{Ij}, \quad j = 1, 2, \ldots, J$$

2. There is a single population of interest, with each individual in the population categorized with respect to two different factors. There are *I* categories associated with the first factor and *J* categories associated with the second factor. For example, Mathematics majors can be classified according to both year [freshman, sophomore, junior, senior, fifth year] and concentration [traditional, secondary ed, statistics, applied].

In type 2 situations, we investigate whether the categories of the two factors occur independently of one another in the population. The null hypothesis reflects the test for independence of joint random variables introduced in Chapter 5.

$$H_0 : p_{ij} = p_{i\cdot} * p_{\cdot j}, \quad i = 1, 2, \ldots, I;\ \ j = 1, 2, \ldots, J$$

$$\text{where}\ \ p_{i\cdot} = \sum_j p_{ij}\ \text{and}\ p_{\cdot j} = \sum_i p_{ij}$$

The test statistic used in testing for homogeneity is identical to the one used in testing for independence. We're going to illustrate the process in Example A.

The determination of degrees of freedom is the same for both cases, as well. This is because the number of freely determined cell counts is *IJ* – 1, since only the total *n* is fixed in advance. Also, there are *I* estimated $p_{i\cdot}$'s, but only *I* – 1 are independently estimated since $\sum p_{i\cdot} = 1$. Similarly *J* – 1 $p_{\cdot j}$'s are independently estimated, so *I* + *J* – 2 parameters are independently estimated. Thus,

$$\text{degrees of freedom } = (IJ-1)-(I+J-2) = IJ - I - J + 1 = (I-1)(J-1).$$

Example A: The Pew Research issued a report in 2012 investigating trends in political party affiliation.
Source: http://www.people-press.org/2012/06/04/section-9-trends-in-party-affiliation/

Generation	Democrat	Independent	Other	Republican
Silent (b. 1928-1945)	34%	27%	5%	34%
Boomer (b. 1946-1964)	34%	34%	5%	27%
Gen X (b. 1965-1980)	29%	42%	5%	24%
Millenial (b. 1981-1994)	31%	45%	6%	18%

Approximate sizes of each population in the United States, based on the 2010 census:
Source: http://www.catalyst.org/knowledge/generations-workplace-united-states-canada

Generation	Size (in millions)
Silent (b. 1928-1945)	40.3
Boomer (b. 1946-1964)	81.5
Gen X (b. 1965-1980)	61.0
Millenial (b. 1981-1994)	85.4

Does the data suggest that the proportions falling in the various political affiliation categories are not the same for the four age groups?

hypotheses:

observed counts: $n_{ij} = p_{ij} * n_i$

Generation	Democrat	Independent	Other	Republican	total
Silent (b. 1928-1945)					
Boomer (b. 1946-1964)					
Gen X (b. 1965-1980)					
Millenial (b. 1981-1994)					
total					

estimated counts (maximum likelihood estimator): $\hat{e}_{ij} = n_i * \frac{n_{\cdot j}}{n} = \frac{(i\text{th row total})(j\text{th column total})}{n}$

Generation	Democrat	Independent	Other	Republican
Silent (b. 1928-1945)				
Boomer (b. 1946-1964)				
Gen X (b. 1965-1980)				
Millenial (b. 1981-1994)				

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \sum_{\text{all cells}} \frac{(\text{observed} - \text{estimated expected})}{\text{estimated expected}}:$$

Generation	Democrat	Independent	Other	Republican
Silent (b. 1928-1945)				
Boomer (b. 1946-1964)				
Gen X (b. 1965-1980)				
Millenial (b. 1981-1994)				

degrees of freedom:

critical value:

conclusion: