Tests of Homogeneity and Independence

Lecture 43

Sections 14.4 - 14.5

Robb T. Koether

Hampden-Sydney College

Wed, Apr 9, 2008

Outline

Introduction

- We wish to compare two or more populations by observing the values of a qualitative variable for each of them.
- For example, in the populations of Democrats, Republicans, and Independents, do they exhibit a different distribution of opinions about Mayor Wilder?
- In this case, we are asking whether the populations are homogeneous.

Introduction

- Or we could say that we wish to investigate the relationship between two or more qualitative variables for a single population.
- For example, given the population of all adults, is one's opinion of Mayor Wilder independent of one's political affiliation?
- In this case, we are asking whether the variables are independent.
- The two questions are the same. The same test will answer both of them.

Homogeneous Populations

Definition (Homogeneous)

Two populations are called homogeneous if they exhibit the same distributions over the same categories.

Examples

- If two colleges' student bodies are each 55% female and 45% male, then the populations are homogeneous.
- If a teacher teaches two sections of Statistics using two different teaching methods and the two sections have the same grade distributions, then they are homogeneous.

Example

- Suppose a teacher teaches two sections of Statistics and uses two different teaching methods.
- At the end of the semester, he gives both sections the same final exam and he compares the grade distributions.
- He wants to know if the differences that he observes are significant.

Example

- Does there appear to be a difference?
- Or are the two populations homogeneous?

	A	B	C	D	F
Method I	5	7	36	17	7
Method II	7	11	18	7	5

The Test of Homogeneity

- The null hypothesis is that the populations are homogeneous.
- The alternative hypothesis is that the populations are not homogeneous.
  - H_0: The populations are homogeneous.
  - H_1: The populations are not homogeneous.

Independence

Definition (Independent)

Two variables are independent if the value of one has no bearing (no predictive value) on the other.

Mendel's Experiments

- In Mendel's experiments, Mendel observed
  - 75% yellow seeds, 25% green seeds.
  - 75% smooth seeds, 25% wrinkled seeds.
- Because color and texture were independent, he also observed
  - $9/16$ yellow and smooth
  - $3/16$ yellow and wrinkled
  - $3/16$ green and smooth
  - $1/16$ green and wrinkled

Mendel's Experiments

- That is, he observed the same ratios within categories that he observed for the totals.

	Smooth	Wrinkled	Ratio
Yellow	9	3	
Green	3	1	
Ratio			

Mendel's Experiments

- That is, he observed the same ratios within categories that he observed for the totals.

	Smooth	Wrinkled	Ratio
Yellow	9	3	3 : 1
Green	3	1	3 : 1
Ratio			

Mendel's Experiments

- That is, he observed the same ratios within categories that he observed for the totals.

	Smooth	Wrinkled	Ratio
Yellow	9	3	
Green	3	1	
Ratio	3 : 1	3 : 1	

Mendel's Experiments

- Had the traits not been independent, he might have observed something different.

	Smooth	Wrinkled	Ratio
Yellow	8	2	
Green	4	2	
Ratio			

Mendel's Experiments

- Had the traits not been independent, he might have observed something different.

	Smooth	Wrinkled	Ratio
Yellow	8	2	4 : 1
Green	4	2	2 : 1
Ratio	2 : 1	1 : 1	

Example

Article

City Hall turmoil: Richmond Times-Dispatch poll

- Observe each person's opinion and political affiliation.
- Do the two variables appear to be independent?

	Excellent	Good	Fair	Poor	Not Sure
Dem	38	70	59	86	16
Rep	21	31	33	23	7
Ind	13	23	31	41	7

The Test of Independence

- The null hypothesis is that the variables are independent.
- The alternative hypothesis is that the variables are not independent.
  - H_0: The variables are independent.
  - H_1: The variables are not independent.

An Example

- We will work the example of the two teaching methods.

	A	B	C	D	F
Method I	5	7	36	17	7
Method II	7	11	18	7	5

The Test Statistic

- The test statistic is the chi-square statistic, computed as

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}.$$

- The question now is, how do we compute the expected counts?

Expected Counts

- Under the assumption of homogeneity or independence (H_0), the rows should exhibit the same proportions.
- We can get the best estimate of those proportions by pooling the rows.
- That is, find the column totals then compute the column proportions from them.

Row and Column Proportions

	A	B	C	D	F
Method I	5	7	36	17	7
Method II	7	11	18	7	5
Col Total					
Percent					

Row and Column Proportions

	A	B	C	D	F
Method I	5	7	36	17	7
Method II	7	11	18	7	5
Col Total	12	18	54	24	12
Percent					

Row and Column Proportions

	A	B	C	D	F
Method I	5	7	36	17	7
Method II	7	11	18	7	5
Col Total	12	18	54	24	12
Percent	10%	15%	45%	20%	10%

Expected Counts

- Similarly, the columns should exhibit the same proportions, so we can get the best estimate by pooling the columns.
- That is, find the row totals and then compute the row proportions from them.

Row and Column Proportions

	A	B	C	D	F	R T	%
Method I	5	7	36	17	7		
Method II	7	11	18	7	5		
Col Total	12	18	54	24	12		
Percent	10%	15%	45%	20%	10%		

Row and Column Proportions

	A	B	C	D	F	R T	%
Method I	5	7	36	17	7	72	
Method II	7	11	18	7	5	48	
Col Total	12	18	54	24	12		
Percent	10%	15%	45%	20%	10%		

Row and Column Proportions

	A	B	C	D	F	R T	%
Method I	5	7	36	17	7	72	60%
Method II	7	11	18	7	5	48	40%
Col Total	12	18	54	24	12		
Percent	10%	15%	45%	20%	10%		

Row and Column Proportions

	A	B	C	D	F	R T	%
Method I	5	7	36	17	7	72	60%
Method II	7	11	18	7	5	48	40%
Col Total	12	18	54	24	12	120	
Percent	10%	15%	45%	20%	10%		

Row and Column Totals

	A	B	C	D	F	R T
Method I	5	7	36	17	7	72
Method II	7	11	18	7	5	48
Col Total	12	18	54	24	12	120

Expected Counts

- Now apply the appropriate row and column proportions to each cell to get the expected count.
- For example, in the upper-left cell, according to the row and column proportions, it should contain

$$60\% \text{ of } 10\% \text{ of } 120.$$

- That is, the expected count is

$$0.60 \times 0.10 \times 120 = 7.2.$$

Expected Counts

- This is the same as

$$\left(\frac{72}{120}\right) \times \left(\frac{12}{120}\right) \times 120 = \frac{72 \times 12}{120} = 7.2.$$

- Therefore, the quick formula is

$$\text{Expected Count} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}.$$

Expected Counts

- Apply that formula to each cell to find the expected counts and add them to the table.

	A	B	C	D	F
Method I	5 (7.2)	7 (10.8)	36 (32.4)	17 (14.4)	7 (7.2)
Method II	7 (4.8)	11 (7.2)	18 (21.6)	7 (9.6)	5 (4.8)

The Test Statistic

- Now compute χ^2 in the usual way.

$$\begin{aligned}\chi^2 &= \frac{(5-7.2)^2}{7.2} + \frac{(7-10.8)^2}{10.8} + \frac{(36-32.4)^2}{32.4} \\ &\quad + \frac{(17-14.4)^2}{14.4} + \frac{(7-7.2)^2}{7.2} + \frac{(7-4.8)^2}{4.8} \\ &\quad + \frac{(11-7.2)^2}{7.2} + \frac{(18-21.6)^2}{21.6} + \frac{(7-9.6)^2}{9.6} \\ &\quad + \frac{(5-4.8)^2}{4.8} \\ &= 7.2106.\end{aligned}$$

Degrees of Freedom

- The number of degrees of freedom is

$$df = (\text{No. of rows} - 1) \times (\text{No. of cols} - 1).$$

- In our example, $df = (2-1) \times (5-1) = 4$.
- To find the p-value, calculate

$$\chi^2 = \chi^2\texttt{cdf(7.2106,E99,4)} = 0.1252.$$

- At the 5% level of significance, the differences are not statistically significant.

Summary

Summary

- Homogeneity of populations and independence of variables are two different ways of looking at the same thing.
- The χ^2 test is used to settle the question of homogeneity or independence.
- The test statistic is the same as it was for the goodness-of-fit test:

$$\chi^2 = \sum_{\text{all cells}} \frac{(O - E)^2}{E}.$$

Summary

Summary

- The main difference is how the expected counts are found:

$$\text{Expected count} = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Grand total}}.$$