XVI. One Way Tables

Scenario. You will have a sample of size N which has been divided in categories (i.e., you will have attribute or nominal data on your sample). The categories will need to be collectively exhaustive and mutually exclusive.

So far when we have considered attribute data we have considered only two possible outcomes for our experiment: "success" and "failure." Every member of our population must fall into one of these two categories. One way tables will permit more than two classifications for the data.

Data and parameters. You will have a set of *observed* outcomes for your data, ususally expressed in a frequency table:

Outcomes				
	Attribute 1	Attribute 2	$\cdots$	Attribute n
Freq.	o_1	o_2	$\cdots$	o_n

You will know the overall sample size N (which is just the sum of outcomes for each label).

Research Objective. You want to compare the *observed* frequencies with some set of *expected* frequencies. The expected frequencies might come from:

○ Census data on the underlying population. In this case, you are likely trying to decide if your sample is biased in some way. In particular, you would be testing:

$$H_0 : \text{ sample is representative of population}$$
$$H_A : \text{ sample is biased}$$

Under ideal conditions you would *control* for bias of this type by building a stratified random sample. When this is not possible, you can use one way tables to

decide *after the fact* whether or not your sample is biased relative to important strata.

○ A set of theoretical proportions. For example, the categories might be preferences for three different brands of cigarettes determined in a blind trial. If you believe that people really can *not* tell the differences between brands, then the idealized proportions would each be one third. In this case you would be testing:

$$H_0 : \text{experimental proportions are all one thirdl}$$
$$H_A : \text{experimental proportions are not all one third}$$

──────── | Solution Template | ────────

Step 1. If the data is not already presented in tabular form, do so now.

	Outcomes			
	Label 1	Label 2	$\cdots$	Label n
Freq.	o_1	o_2	$\cdots$	o_n

Step 2. Write down the *expected* proportions for each outcome:

Expected Proprotions				
	Label 1	Label 2	$\cdots$	Label n
Expected Prop.	p_1	p_2	$\cdots$	p_n

Step 3. Write down the *expected* frequencies.

Expected Frequencies				
	Label 1	Label 2	$\cdots$	Label n
Expected Freq.	$e_1 = p_1 N$	$e_2 = p_2 N$	$\cdots$	$e_n = p_n N$

where N is the overall sample size.

Step 4. Compute the test statistic:

$$\text{test statistic} = \sum \frac{(\text{Observation} - \text{Expectation})^2}{\text{Expectation}}$$

Step 5. Find the degrees of freedom:

$$\text{degrees of freedom} = \# \text{ of categories} - 1$$

Step 6. Find the cutoff in the chi squared table (Table A-4, page 666 of the text).

Step 7. *Decision Rule.* If the value of the test statistic is *larger than* the cutoff, then *reject* the null hypothesis and accept the alternative. Otherwise, accept the null hypothesis.

——————— *End of Solution Template* ———————

Example. A random sample of 483 students is taken from the OU student body. It is found that this sample has the following ethnic breakdown:

		White	Black	Hispanic	Asian	Indian	Other
				Observations			
#		386	24	27	12	20	14
%		79.9	5.0	5.6	2.5	4.1	2.9

It is known that the ethnic breakdown of the OU student body is:

White	Black	Hispanic	Asian	Indian	Other
75.8%	.061	2.3%	3.5%	4.7%	7.6%

Does the sample differ significantly ($\alpha = 0.5\%$) from the overall student population with respect to ethnicity?

Solution.

Step 1. The data is already presented in appropriate form.

Step 2. We have been given expected *percentages*; we need expected *proportions*, so the table is

	White	Black	Hispanic	Asian	Indian	Other
			Expected Proportions			
	.758	.061	0.0 23	0.035	0.047	0.076

$\boxed{\textit{Step 3.}}$ The expected frequencies table looks like:

Expectations						
	White	Black	Hispanic	Asian	Indian	Other
#						

We need to fill in the blanks with what we would expect our sample to look like if it were 100% representative of the OU student body (no sampling error). Thus, for example, we would expect 75.8% of the total sample to be white:

$$\text{expected \# of whites} = 0.758 \times 483 = 366.11$$

giving the first entry in the table. Proceeding in this fashion gives:

Expectations						
	White	Black	Hisp.	Asian	Indian	Other
#	366.11	29.46	11.11	16.91	22.70	36.71

$\boxed{\textit{Step 4.}}$ The value of the test statistic is

$$\chi^2 = \sum \frac{(\text{Observations} - \text{Expectations})^2}{\text{Expectations}}$$
$$= \frac{(386 - 366.11)^2}{366.11} + \frac{(24 - 29.46)^2}{29.46} + \cdots$$
$$\cdots + \frac{(27 - 11.11)^2}{11.11} + \frac{(12 - 16.91)^2}{16.19} + \cdots$$
$$\cdots + \frac{(20 - 22.70)^2}{22.70} + \frac{(14 - 36.71)^2}{36.71}$$
$$= 1.08 + 1.01 + 22.73 + 1.43 + 0.32 + 14.05$$
$$= 40.62$$

$\boxed{\textit{Step 5.}}$ Since there are six categories, there are five degrees of freedom.

$\boxed{\textit{Step 6.}}$ The cutoff from the chi-squared table is thus 16.75.

$\boxed{\textit{Step 7.}}$ For these kinds of problems, chi-squared tests are *one-sided*, so we reject the null hypothesis if the test statistic is *greater than the cutoff.* In this case, the

test statistic 40.62 which is much larger than the cutoff of 16.75, so we reject the null hypothesis and accept the alternative, namely that the sample is not representative of the general population with respect to ethnicity. ∎

Remark. When you have only two categories ("success" and "failure") it is possible to perform one sided tests as well as the two sided tests discussed here. The details, similar in outline to what we have done in this secion, are discussed in your textbook.

Problems

1. Given the following observed and expected values, deterimine if there is a significant difference between the two (use a significance level of 5%):

observed	45	32	38	51
expected	25%	25%	25%	25%

2. Given the following observed and expected values, deterimine if there is a significant difference between the two (use a significance level of 1%):

observed	654	598	621
expected	38%	31%	31%

3. Given the following observed and expected values, deterimine if there is a significant difference between the two (use a significance level of 1%):

observed	34	38	12	8
expected	30	40	15	15