χ^2 AS A TEST FOR HYPOTHESES (INDEPENDENCE OF VARIABLES)

A. A. Introduction

1. the distribution, the chi-square, can be used to test "goodness of fit" or to test the independence of variables

2. we will use it to test the independence of variables

Introduction (cont’d)

3. chi-square is a **non-parametric** test for/of hypotheses

a. it is non-parametric because it does not rely upon two assumptions that parametric inferential tests need, i.e.,

(i) that the original distribution is normal

(ii) that the analyst can use the Central Limit Theorem

Introduction (cont'd)

b. non-parametric tests do not rely on the estimate of parameters, although the χ^2 involves the median when used to test goodness of fit

c. the lack of parametric assumptions casts some doubt on the ability of a χ^2 analysis to generalize from one sample to a population – there is some serious contention about this point

Introduction (cont'd)

4. we will use the to test hypotheses when the data are categorical, i.e., when measured using two characteristics, e.g., height and weight, years of computer experience and hours of Web use daily, and gender and political preference, they fall into mutually exclusive and collectively exhaustive categories

5. the data are then organized into cells in what is called a contingency table or cross break

6. the number of cells is equal to the number of rows multiplied by the number of columns (R x C)

B. An extended example – generating the contingency table

1. our example comes from Busha & Harter (1980, pp. 301-305)

2. the data below are from a study of gender and political preference as expressed by political party affiliation

3. the null hypothesis -- H_0 : There is no relationship between political preference and gender at $\alpha = 0.05$.

Generating the contingency table (cont'd)

4. let's use the observations to generate a contingency table or cross break, with marginals (also called marginal totals)

Generating the contingency table (cont'd)

5. every cross break is an R x C table, in this case, a 4 x 2 contingency table; generally, matrices are described as R x C

6. we then have to determine E for each cell, the expected value in each cell if the null hypothesis were true, i.e., if the proportion by gender of each political preference were the same as the proportion by gender for each row and each column.

Generating the contingency table (cont'd)

a. another way of saying this is that expected values are what we would find in each cell if the total proportion of women to men were to hold true for each party affiliation

b. $E = \frac{T_R x T_c}{n}$; for our example, I'll do two or three cells

c.

55.4	100.6
27.3	49.7
14.2	25.8
13.1	23.9

Generating the contingency table (cont'd)

d. d. recall that every E answers the question: what values would be in each cell if (the null hypothesis) were true?

7. ∃ differences between the observed values (O) and the expected values (E)

a. are the differences due to chance alone (random variability or noise), or are they statistically significant at this α ?

Generating the contingency table (cont'd)

b. are gender and political preference independent (we would then retain H_0), or are they related (reject H_0) at this α ?

c. given these sample data, must we reject or fail to reject H_0 at this α ?

d. the χ^2 test of independence, at a particular α , helps us to determine if the differences between E and O are statistically significant and whether we should reject or fail to reject (retain) this null hypothesis

C. Testing the null hypothesis using χ^2

1. recall the fundamentals of hypothesis testing

   Refer to Hypothesis Testing decision tree

Testing the null hypothesis using χ^2 (cont'd)

2. one way of thinking about χ^2 is as a measure of the squared distance or difference between observed and expected values in a distribution of categorical data

a. remember that, the deviation score we used in determining z-scores, was the deviation between x and μ ; with χ^2 we are measuring the squared variation between E and O

Testing the null hypothesis using χ^2 (cont'd)

b. the particular calculated values of chi square:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

c. as the differences between O and E increase, so does

χ^2

Testing the null hypothesis using χ^2 (cont'd)

d. a brief aside

(i) if E (the denominator) is small, we may get a large value for χ^2 even if the differences between E and O are relatively small

(ii) if E (the denominator) is large, we may get a small value for χ^2 even if the differences between E and O are relatively large

(iii) $\therefore$ the expected frequency in each cell must ≥ 5; if not, then cells must be combined, thereby reducing the (R-1)(C-1) degrees of freedom by the number of cells combined with others

Testing the null hypothesis using χ^2 (cont'd)

3. shape of the χ^2

   Refer to Spatz (1997, p. 290, Figure 12.1)

Testing the null hypothesis using χ^2 (cont'd)

4. χ^2 table

Refer to Spatz (1997, p. 372, Appendix B, Table E

Testing the null hypothesis using χ^2 (cont'd)

a. compare the table value of χ^2 to the calculated value of χ^2

b. df = (R-1)(C-1)

where R = the number of rows and C = the number of columns

c. α = the significance level

Testing the null hypothesis using χ^2 (cont'd)

5. let's calculate the χ^2 of our example – remember that α = 0.05 and that the number of terms in the summation will equal the number of cells in the cross break (R x C), in this case, eight (8)

$$\chi^2 = \sum \frac{(O-E)^2}{E} + \frac{(62-55.4)^2}{55.4} + \frac{(35-27.3)^2}{27.3} + \frac{(7-14.2)^2}{14.2} + \frac{(6-13.1)^2}{13.1}$$

$$+ \frac{(94-100.6)^2}{100.6} + \frac{(42-49.7)^2}{49.7} = \frac{(33-25.8)^2}{25.8} + \frac{(31-23.9)^2}{23.9}$$

= 0.79 + 2.17 + 3.65 + 3.85 + 0.43 + 1.19 + 2.01 + 2.11

= 16.2

Testing the null hypothesis using χ^2 (cont'd)

6. determine χ^2 from the table for this contingency table and significance level

a. df = (R-1)(C-1) = (4-1)(2-1) = 3

b. α = 0.05

Testing the null hypothesis using χ^2 (cont'd)

c. we can use the convention χ^2 (df, α)

(i) χ^2 (3, 0.05) in our example

(ii) χ^2 (3, 0.05) = 7.82

Testing the null hypothesis using χ^2 (cont'd)

7. compare $\chi^2 calc$ to $\chi^2 table$

a. if $\chi^2 calc \leq \chi^2 table$

(i) data are likely to have occurred by chance; differences between E and O are likely only noise

(ii) data are not significant

(iii) do **not** reject the null (H_0)

(iv) gender and political preference are independent

Testing the null hypothesis using χ^2 (cont'd)

b. if $\chi^2 calc > \chi^2 table$

(i) data are **not likely** to have occurred by chance; differences between E and O are likely to have occurred because of some (as yet unspecified) systematic relationship between gender and political preference

(ii) the data are significant

(iii) reject the null (H_0)

(iv) gender and political preference are related

Testing the null hypothesis using χ^2 (cont'd)

c. in our example

(i) $\chi^2 table = 7.82$, $\chi^2 calc = 16.2$

(ii) 16.2>>7.82 (16.2 is much greater than 7.82)

Testing the null hypothesis using χ^2 (cont'd)

d. therefore, we reject H_0 and conclude that gender has a statistically significant effect on political preference

(i) determining what effect(s) might be involved is a matter of looking at the data closely, especially noting those cells where the differences between E and O are greatest

(ii) as noted above, how far one can generalize from a sample to the population using χ^2 is a matter of some controversy; remember that this controversy springs largely from the fact that χ^2 is a non-parametric test

D. Reprise

1. let's review the fundamentals of hypothesis testing

Refer to Hypothesis Testing Decision Tree

Reprise (cont'd)

2. we used χ^2, a non-parametric test, to test a statistical hypothesis formally

a. we reached a conclusion that it was not likely that the results occurred by chance alone

b. we then used the results of the statistical test to reject or fail to reject the null hypothesis

Reprise (cont'd)

3. reconsider some themes from our discussion of the *Literary Digest* U.S. presidential poll of 1936

a. that poll, like the others that preceded it, had a grossly unrepresentative sample – a sample that was disproportionately richer, older, more male, "whiter," and more Republican than the voting population

b. one could still be right with such a biased (systematic error in one direction), unrepresentative sample – but such an outcome is highly unlikely

Reprise (cont'd)

c. similarly, one could be wrong in one's conclusions even with a highly representative sample – but such an outcome would also be unlikely

d. we can always be wrong when making a probabilistic argument – but we use inferential statistics to tell us how likely if is that we are wrong in the long run

e. always remember Type I (α) and Type II (β) error