PART III CHI-SQUARE

Topic 1. Introduction To The Concept of Statistical Tests Using Chi-Square

Assumptions

Assume 100 customers tried each of three products and selected the one they like best. (The 100 represent a population of interest to us.)

Product	1	2	3
People Preferring	25	35	40

Two Different Scenarios

Scenario 1. Now assume 50 of the customers were male, and 50 were female.

- Now suppose we cross-classified by Gender and Preference.

Table 1 Product Preferences

		P1	P2	P3	
Gender	Male				50
	Female				50
		25	35	40	100

- Questions to think about:

1) What would this table look like if there was no relationship between the variable Gender and the variable Product Preference?

2) What would the table look like if there was a relationship between the variable Gender and the variable Product Preference?

Scenario 2. Now assume 60 of the customers were male, and 40 were female

- Now suppose we cross-classified by Gender and Preference.

Table 2 Product Preferences

		P1	P2	P3	
Gender	Male				60
	Female				40
		25	35	40	100

PART III CHI-SQUARE

Topic 1. The Concept (Continued)

- Questions to think about:

  1) What would this table look like if there was no relationship between the variable Gender and the variable Product Preference?

  2) What would the table look like if there was a relationship between the variable Gender and the variable Product Preference?

Note

1. There is only one way the table can look in the case where there is no relationship between the variables; but there are a very large number of tables that could result in the event that there was a relationship between the variables.

2. There is no relationship between two variables when the proportion of responses falling in a particular row of every column is the same as the proportion of responses falling in that row for the table as a whole.

Analysis

Now suppose when we cross-classified the data for Gender and Product Preference, we observed the following table:

Table 3

		Product Preferences			
		P1	P2	P3	
Gender	Male	6	26	28	60
	Female	19	9	12	40
		25	35	40	100

Do you believe this table suggests there is a relationship between the two variables or not?

Suppose our goal is to construct a carefully defined index such that-- the larger the index,

the larger the difference between what we actually observed and what we would have expected to observe if there were no relationship.

PART III CHI-SQUARE

Topic 1. The Concept (Continued)

Now, we need to construct what **table 3** would have looked like if there were **no relationship** between the two variables.

$\frac{X}{25} = \frac{60}{100}$, 100X = 150, X = 15 $\qquad$ $\frac{X}{35} = \frac{60}{100}$, 100X = 2100, X = 21

$\frac{X}{40} = \frac{60}{100}$, 100X = 2400, X = 24 $\qquad$ $\frac{X}{25} = \frac{40}{100}$, 100X = 1000, X = 10

$\frac{X}{35} = \frac{40}{100}$, 100X = 1400, X = 14 $\qquad$ $\frac{X}{40} = \frac{40}{100}$, 100X = 1600, X = 16

Table 4 Product Preferences

		P1	P2	P3	
Gender	Male	15	21	24	60
	Female	10	14	16	40
		25	35	40	100

Note that the proportion of responses in the row of each column is precisely the same as the proportion of responses in that row of the table as a whole.

Chi-square Definition

$$X^2 = \sum_{i=1}^{n} \left[\frac{(f_o^i - f_e^i)^2}{f_e^i} \right]$$

- Using observations in **Table 3**, we calculate Chi-square as:

$$X^2 = \frac{(6-15)^2}{15} + \frac{(26-21)^2}{21} + \frac{(28-24)^2}{24} + \frac{(19-10)^2}{10} + \frac{(9-14)^2}{14} + \frac{(12-16)^2}{16}$$

$$= \frac{81}{15} + \frac{25}{21} + \frac{16}{24} + \frac{81}{10} + \frac{25}{14} + \frac{16}{16}$$

$$= 5.4 + 1.19 + .67 + 8.1 + 1.79 + 1 = 18.15$$

PART III CHI-SQUARE

Topic 1. The Concept (Continued)

- Now suppose the table we actually observed was as follows:

Table 5 Product Preferences

		P1	P2	P3	
Sex	Male	14	20	26	60
	Female	11	15	14	40
		25	35	40	100

$$X^2 = \frac{(14-15)^2}{15} + \frac{(20-21)^2}{21} + \frac{(26-24)^2}{24} + \frac{(11-10)^2}{10} + \frac{(15-14)^2}{14} + \frac{(14-16)^2}{16}$$

$$= \frac{1}{15} + \frac{1}{21} + \frac{4}{24} + \frac{1}{10} + \frac{1}{14} + \frac{4}{16}$$

$$= .067 + .048 + .167 + .1 + .071 + .25 = .703$$

- Note also, If there was no relationship between the variables, Chi-square would be exactly 0.

Sampling Error Concern

1. Possible problems

   Up till now we have not concerned ourselves with the question of sampling error. We have treated the tables as if they were (or were perfectly representative of) the population.

   But now suppose the 100 people in **Table 3** were a 10% random sample from a population of 1,000.

   Could **Table 3** have been a random sample from a population that was distributed as **Table 6**?

Table 6 Product Preferences

		P1	P2	P3	
Gender	Male	150	210	240	600
	Female	100	140	160	400
		250	350	400	1000

Topic 1. The Concept (Continued)

Although it is unlikely, **Table 3** could be a random sample from **Table 6**.

This example demonstrates that it is possible, because of sampling error, to obtain a sample in which there "appears to be a relationship between two variables" from a population where no such relationship exists.

2. Solution to sampling error problem

So, we must devise a method for assessing the likelihood (probability) that a particular sample came from a population where no relationship exists between the variables. The solution to this problem, although tedious, is logically simple.

- Suppose we have an actual random sample of size n from a population of size N.
- We first calculate the X^2 value for the sample.
- Now we wish to know the probability that we could have obtained our sample X^2 from a population where there is no relationship between the variables --from a population whose true X^2 was really zero.

  To obtain this probability takes several steps.

  1) First, we assume a population cross-classification table with the same number of rows and columns as our sample but which has no relationship between the variables--that is, a population whose true X^2 is equal to zero.

  2) Then we take every possible sample (of size n) from this population, and, for each sample, calculate the sample's X^2. Note that for any reasonably sized population (N) and sample (n) this will result in a very large number of sample X^2 values.

  3) Now we arrange this "sampling distribution" into a frequency distribution which we convert to a probability distribution.

  4) Now, we know that this probability distribution of X^2 values came from a population whose true X^2 is zero.

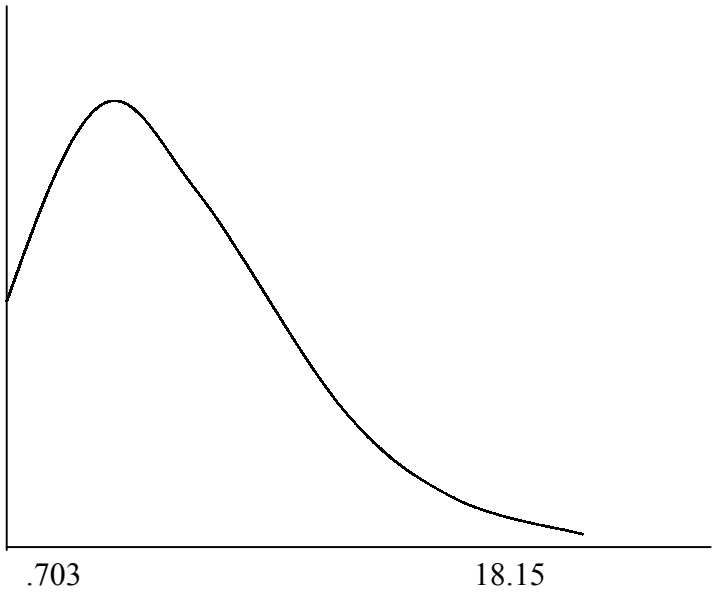
  5) Then, for any given value of X^2, we can calculate <u>the probability that we will obtain a X^2 of that value or larger, given that the population X^2 is really zero.</u>

PART III CHI-SQUARE

Topic 1. The Concept (Continued)

Figure 1

Probability distribution of sample X^2 values from a population whose true X^2 value is zero.



3. Calculating the appropriate probability for our example

- Now we are able to assess the probability that the X^2 value from our observed data in **Table 4** could have come from a population whose true X^2 value was zero.

    1) The area to the right of 18.15 in the probability distribution shown in **Figure 1** is a good deal less than .001. Prob ($X^2(2) > 18.15$) = .00012

    2) Since we know the probability distribution was derived from a population whose true X^2 was 0, there is a .012% chance of getting a X^2 value greater than or equal to 18.15 from a random sample taken from a population whose true X^2 is equal to 0.

    3) Thus, there is less than a .1% chance of observing our sample X^2 of 18.15 or larger if the population X^2 the sample came from is zero. (It can happen, but the odds are very low--less than 1 in one thousand.)

    4) Given such a low probability, we must reject the assumption that our sample came from a population whose true X^2 was zero and conclude that the population's X^2 is not zero.

    5) Therefore, there is probably a relationship between the variables in the population.

PART III CHI-SQUARE

Topic 1. The Concept (Continued)

To recap the logic:

If it is very unlikely that our sample X^2 came from a population whose true X^2 is zero, then we conclude that our random sample came from a population whose true X^2 is not zero--that our sample came from a population where there is a relationship between the variables.

- In contrast, consider the sample X^2 value from **Table 5**.
  1) About 70% of the time, we will get a sample which yields a X^2 of .703 or larger from a population whose true X^2 is zero.

  2) Since the odds are not unusually small, we cannot reject the assumption that our sample came from a population whose true X^2 is zero. And we conclude it is entirely possible for our sample to have come from a population whose true X^2 is zero and that there is probably no relationship between the variables in the population.

Most texts give you "critical probabilities" which is a euphemism for probabilities the authors can live with. In contrast, I want you to understand the logic underlying the process so well that you can do without the crutch of someone else's critical probabilities. Rather, I want you to be able to argue for or against a relationship between variables in a population because you have a perfect understanding of the logic underlying your ability to reach such a conclusion.

Topic 2. Recapping The Logic

1. We are <u>not</u> trying to see if our sample is representative of the population – we must simply assume that it is.

2. We are trying to decide whether our sample came from a population where:
   a) Chi-square is equal to 0;
   Or b) Chi-square is not equal to 0.

3. If the probability that one would get a Chi-square as large or larger than our sample's Chi-Square value is low (less than our cutoff), then we reject "Chi-square equals 0 in the population." And if Chi-Square in the population is not 0, then we conclude a relationship must exist between the variables in the population where this sample came from.

4. But if our probability is above the cutoff, then we cannot reject "Chi-Square equal 0 in the population." And if Chi-Square in the population is 0, then we conclude a relationship does not exist between the variables in the population where this sample came from.

PART III CHI-SQUARE

Topic 3. Steps In Conducting A Chi-Square Test

1. Decide on your cutoff probability.

2. Do any necessary recodes
   a) natural breaks
   b) equal intervals
   c) equal # of observations

3. Run your table

4. Check that all the "expects" are greater than or equal to 1 and 80% are greater than or equal to 5.

5. Look at probability and Chi-square

6. Reach conclusion about whether you think:
   i) a relationship exists in the population
   ii) a relationship does not exist in the population

7. If ii), then stop (Really Stop);
   If i), then interpret table and other measures to be discussed

8. State your overall business conclusion

For Presentation

a) 1 page with 2 histograms and breaks noted
b) 1 page with Chi-Square table (# observed, # expected, col %, row %)
c) optional graph of results