**GPPI PPOL 501–02 & -06: Fall 2014**
**Course Notes # 15: Chi-Square Tests**
**Professor Carolyn Hill**

---

## I. CHI-SQUARE ($\chi^2$) TESTS

***Two Types***

- The Chi-Square test can also be used to examine the **distributions** of two similarly-defined variables (e.g., is the distribution of educational attainment the same for prisoners in 1980 as it is in 1998?)

  This use of the Chi-Square test is called the Chi-Square Goodness-of-Fit Test

- Chi-Square tests can be used to examine **association** or **independence** between independent and dependent variables of interest (e.g., calendar year and whether parents help students with homework; between student gender and feelings of safety)

  This use of the Chi-Square test is called the Chi-Square Test of Independence.

***Key Points:***

- Chi-Square tests can be used for nominal, ordinal, or interval-ratio variables (most often for nominal and ordinal variables, however).

- After an example, we'll talk about the advantages and limitations further.

## II. BASIC PROPERTIES OF THE $\chi^2$ CURVE

- total area under the $\chi^2$ curve is equal to 1.

- A $\chi^2$ curve starts at 0 on the horizontal axis and is right-skewed.

- The curve never touches the horizontal axis.

- As the number of d.f. becomes larger, the $\chi^2$ curves look increasingly like Normal curves

## III. CHI-SQUARE TEST: GOODNESS-0F-FIT APPLICATION

- One way that Chi-Square tests can be used is to examine the **distributions** of two similarly-defined variables.

  This use of the Chi-Square test is called the Chi-Square Goodness-of-Fit Test

  **Example:** In 1980, the distribution of educational attainment of death row inmates was:

| Education | Percent |
|---|---|
| 8th grade or less | 25.7 |
| 9th – 11th grade | 37.0 |
| High school grad / GED | 29.5 |
| Any college | 7.8 |

  In 1998, 128 prisoners who were sentenced to death were randomly selected. Their educational attainment was:

| Education | Frequency |
|---|---|
| 8th grade or less | 18 |
| 9th – 11th grade | 48 |
| High school grad / GED | 49 |
| Any college | 13 |
| TOTAL | 128 |

  Do the data provide sufficient evidence that the educational attainment distribution of prisoners sentenced to death in 1998 differs from that of 1980 death-row inmates?

  H0: The educational attainment distribution of prisoners on death row in 1998 is the same as the 1980 distribution.

  H1: The educational attainment distribution of prisoners on death row in 1998 is different from the 1980 distribution.

- The logic of the Chi-Square test is to compare what we *actually* observe (i.e., the frequencies for each category in the sampled prisoners) with what we would *expect* to observe if the distribution remained unchanged.

$$\chi^2(obtained) = \sum\left(\frac{(f_o - f_e)^2}{f_e}\right)$$

Where $f_o$ = Observed frequency for each cell
$f_e$ = Expected frequency for each cell

$$\chi^2(obtained) = \sum\left(\frac{(f_o - f_e)^2}{f_e}\right)$$

| Cell | Observed Frequency $f_o$ | Expected Frequency $f_e$ | Deviations $(f_o - f_e)$ | Squared Deviations $(f_o - f_e)^2$ | Weighted Squared Deviations $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|---|---|
| 8th grade or less | 18 | 32.9 | -14.9 | 221.9 | 6.7 |
| 9th – 11th grade | 48 | 47.4 | 0.6 | 0.4 | 0.0 |
| High school grad / GED | 49 | 37.8 | 11.2 | 126.3 | 3.3 |
| Any college | 13 | 10.0 | 3.0 | 9.1 | 0.9 |
| **TOTAL** | **128** | **128** | | | **11.01** |

- For the Chi-Square goodness-of-fit test, the d.f. are determined as: **d.f. = k-1,** where *k* is the number of possible categories for the variable being examined.

  In this case, the variable is "educational attainment of death-row inmates" and there are k=4 possible categories. So d.f. = 3.

- $\chi 2$ critical value = 7.81 at 95% confidence level
  = 11.35 at 99% confidence level

- Compare $\chi 2$ calculated to $\chi 2$ critical: 7.81 < 11.01 < 11.35

- Make a decision: Reject the null at the 95% level; Fail to reject at the 99% level

## IV. CHI-SQUARE TESTS: TESTING INDEPENDENCE

- Chi-Square tests also can be used to examine **association** or **independence** between independent and dependent variables of interest (e.g., calendar year and whether parents help students with homework; between student gender and feelings of safety)

  This use of the Chi-Square test is called the Chi-Square Test of Independence.

- This example is drawn from a (roughly) 5% random sample of 8th graders in the Chicago Public Schools (CPS) who responded to a survey produced by the Consortium on Chicago School Research in 1994 and 1997.

- During this period, the district was engaged in a widespread school reform effort that involved a variety of initiatives, including creating local school councils, placing low-performing schools on probation, requiring low-achieving students to attend summer school and focusing on a systemwide safety-discipline concerns.

The data set includes the following variables:

**SAFETY** "How safe do you feel outside around your school?"
Measurement Level: Ordinal

| Value | Label |
|---|---|
| 1 | not safe |
| 2 | somewhat safe |
| 3 | mostly safe |
| 4 | very safe |

**PARHELP** "How often do your parents help you with your homework?"
Measurement Level: Ordinal

| Value | Label |
|---|---|
| 1 | never |
| 2 | once in a while |
| 3 | most of the time |
| 4 | all of the time |

**RACE** **Race**
Measurement Level: Nominal

| Value | Label |
|---|---|
| 1 | White |
| 2 | Black |
| 5 | Hispanic |

**GENDER** **Gender**
Measurement Level: Nominal

| Value | Label |
|---|---|
| 1 | male |
| 2 | female |

In each of the questions below, we'll
(a) state our prior expectations informally;
(b) state the null hypothesis and alternative hypotheses formally;
(c) present a crosstab with statistical results that bear on the question;
(d) state the implications / inferences of our findings (i.e., "reject the null" or "fail to reject the null" and why)

**EXAMPLE 1. Do boys and girls differ in how safe they feel around school?**

First, let's do a simple *crosstabulation* of these variables:

**Table 1: Cross-tabulation of gender and feeling safe around school (actual frequencies only)**

| | | GENDER | | Row Total |
|---|---|---|---|---|
| | | ***male*** | ***female*** | |
| **How safe are you outside around School?** | ***not safe*** | 238 | 257 | 495 |
| | ***somewhat safe*** | 310 | 421 | 731 |
| | ***mostly safe*** | 356 | 419 | 775 |
| | ***very safe*** | 211 | 234 | 445 |
| **Column Total** | | 1115 | 1331 | 2446 |

*Things to note….*

- The "Y" variable, "dependent" variable, variable to be explained, or survey response of interest is usually listed in the rows by convention (here: "How safe are you outside around school?"), and the "X" variable, "independent" variable, or variable that will explain is usually listed in the columns by convention (here, GENDER).

- We usually identify tables by their numbers of rows and columns of possible categories: row by column. This table is a "four by two" table, often abbreviated 4 X 2. (because there are 4 categories of the "Y" and 2 categories of "X")

- Recall that the furthest right column and the bottom row are called "marginal frequencies."

*Now, on to analysis of the association between these two variables…*

(a) Informal expectation: males and females probably don't differ regarding whether they feel safe outside of school (note: you may have another informal expectation)

(b) Formal hypothesis statement:
$H_o$: There *is no association* between gender and feelings of safety outside school. (or, you may say: boys and girls do not feel different levels of safety outside school; or gender and feelings of safety are independent).
$H_1$: There *is an association* between gender and feelings of safety outside school.

(c) Use a Chi-Square test to test this hypothesis. *The logic of the Chi-Square test is:*

- Compare the actual cell frequencies (given in the table above) to a set of hypothetical cell frequencies that that we would expect if there were *no association* between the two variables of interest (i.e., if the null hypothesis were true).

- If the *actual* frequencies are not very different from the *hypothetical* frequencies that we would expect if there were no association, then we conclude that it is likely there *is no* association between the two variables.

- If the *actual* frequencies are very different from the *hypothetical* frequencies that we would expect if there were no association, then we conclude that it is likely there *is* an association between the two variables.

- To assess the meaning of "are not very different" and "are very different," we use a particular type of statistical distribution: the Chi-Square distribution

- The Chi-Square test statistic is: $\chi^2(obtained) = \sum\left(\frac{(f_o - f_e)^2}{f_e}\right)$

- Let's look at an example so you can see how the calculations work:

**Table 2: Cross-tabulation of gender and feeling safe around school (actual and expected freq)**

| | | | gender | | Row Total |
|---|---|---|---|---|---|
| | | | ***male*** | ***female*** | |
| **How safe are you outside around School?** | ***not safe*** | Count | 238 | 257 | 495 |
| | | Expected Count | 225.6 | 269.4 | 495 |
| | | | | | |
| | ***somewhat safe*** | Count | 310 | 421 | 731 |
| | | Expected Count | 333.2 | 397.8 | 731 |
| | | | | | |
| | ***mostly safe*** | Count | 356 | 419 | 775 |
| | | Expected Count | 353.3 | 421.7 | 775 |
| | | | | | |
| | ***very safe*** | Count | 211 | 234 | 445 |
| | | Expected Count | 202.9 | 242.1 | 445 |
| **Column Total** | | | 1115 | 1331 | 2446 |

- Where do the expected frequency counts come from? Look at the expected count (or "hypothetical" count if there is no association between gender and feelings of safety) for the cell of FEMALES who feel "NOT SAFE":

  Expected Frequency = 269.4 $= \left( {}^{495}\!\!\diagup\!\!{}_{2,446} \right) * 1{,}331$

  Basically, this expected count gives a number that would reflect no association between gender and feelings of safety: we would expect the proportion of females who feel "not safe" to be the same as the proportion of all sample members who do not feel safe.

  Or, we could look at it this way: Expected Frequency = 269.4 $= \left( {}^{1331}\!\!\diagup\!\!{}_{2,446} \right) * 495$

  This is another way of looking at the same thing: we expect the proportion of those who feel "not safe" who are female to be the same as the proportion of all sample members who are female.

Calculating the chi-square test:

| Cell | Observed Frequency $f_o$ | Expected Frequency $f_e$ | Deviations $(f_o - f_e)$ | Squared Deviations $(f_o - f_e)^2$ | Weighted Squared Deviations $\frac{(f_o - f_e)^2}{f_e}$ |
|---|---|---|---|---|---|
| males who feel "not safe" | 238 | 225.6 | 12.4 | 153.8 | 0.68 |
| males who feel "somewhat safe" | 310 | 333.2 | -23.2 | 538.2 | 1.62 |
| males who feel "mostly safe" | 356 | 353.3 | 2.7 | 7.3 | 0.02 |
| males who feel "very safe" | 211 | 202.9 | 8.1 | 65.6 | 0.32 |
| females who feel "not safe" | 257 | 269.4 | -12.4 | 153.8 | 0.57 |
| females who feel "somewhat safe" | 421 | 397.8 | 23.2 | 538.2 | 1.35 |
| females who feel "mostly safe" | 419 | 421.7 | -2.7 | 7.3 | 0.02 |
| females who feel "very safe" | 234 | 242.1 | -8.1 | 65.6 | 0.27 |
| **TOTAL** | **2446** | **2446** | **0.0** | **1529.8** | **4.85** |

- So the Chi-Square statistic that we calculate is 4.85.

- We need to compare this to a Chi-Square critical value. To find the appropriate value, we need to identify the appropriate curve, using the degrees of freedom, where

  df = (# rows – 1)* (# columns – 1)

  In this case, df = (4-1)* (2-1) = 3*1 = 3.

- Use the Chi-Square table to find the $\chi^2$ critical value (we go through the same kind of process that we did to test hypotheses with sample means):

- *$\chi$2(3 d.f.) critical value* for a 95% confidence level is = 7.815. (i.e., 5% of the distribution is beyond this value). If the null hypothesis is true (i.e., no association), then 95% of the possible values of the *$\chi$2* statistic would be less than 7.815, and 5% would be greater than 7.815.

- Our *$\chi$2 test statistic* value is 4.85, which is less than 7.815. So,

  (d) We FAIL TO REJECT THE NULL HYPOTHESIS of no association between gender and feelings of safety around school.

---

Try some additional questions for interpretation:

**EXAMPLE 2. Do 8th graders of different race/ethnicities differ in how safe they feel around school?**

(a) State your informal expectations……

(b) State the formal null and alternative hypotheses……

(c) Show the cross tab and calculate the Chi-Square statistic

**Table 3: Race and School Safety**

| | | | Race | | | Row Total |
|---|---|---|---|---|---|---|
| | | | ***White*** | ***Black*** | ***Hispanic*** | |
| **How safe are you outside around School?** | ***not safe*** | Count | 41 | 348 | 106 | **495** |
| | | Expected Count | 79.1 | 279.5 | 136.4 | **495** |
| | ***somewhat safe*** | Count | 94 | 437 | 200 | **731** |
| | | Expected Count | 116.9 | 412.7 | 201.4 | **731** |
| | ***mostly safe*** | Count | 176 | 362 | 237 | **775** |
| | | Expected Count | 123.9 | 437.6 | 213.6 | **775** |
| | ***very safe*** | Count | 80 | 234 | 131 | **445** |
| | | Expected Count | 71.1 | 251.2 | 122.6 | **445** |
| **Column Total** | | | **391** | **1381** | **674** | **2446** |

**Question: What $\chi 2$ critical value should you compare the test statistic to?**

**d.f. = _____**

**$\chi 2$ critical value = _____ at 95% confidence level**

**_____ at 99% confidence level**

**_____ at 99.9% confidence level**

(d) State the conclusion from the statistical test:

## V. ADVANTAGES AND LIMITATIONS OF THE CHI-SQUARE TEST

Advantages

- Can be used with nominal, ordinal, or interval/ratio variables
- Does not require a specific distributional shape for the population (e.g., Normal, uniform, etc.)
- Can be used with variables that have many categories or scores, BUT….

Limitations

- Becomes difficult to interpret when the variables have many categories. Rough rule of thumb: Chi-square tests are easiest to interpret and understand when both variables have 4 or fewer scores
- Sensitive to sample size I: the probability of rejecting the null hypothesis increases as the number of cases increases. (Think more about this in your problem set)
- Sensitive to sample size II: When sample size is small, can't assume that the sampling distribution follows a Chi-Square distribution.

  "Small sample size" interpreted as meaning that a high percentage of cells have expected frequencies of 5 or less. What's a "high percentage of cells"? A conservative strategy is to start to worry if *any cell* has an expected frequency of 5 or less.

  One strategy is to collapse or recombine categories so that no cell has an expected frequency of 5 or less.

*****EXAMPLE OF RECODING TO MEET REQUIREMENTS OF CHI-SQUARE TEST*****

```
. use "J:\STATA datasets\practice.dta", clear

. do "C:\DOCUME~1\gppilab\LOCALS~1\Temp\STD01000000.tmp"

. describe, short

Contains data from J:\STATA datasets\practice.dta
  obs:         2,188
 vars:             5
 size:        26,256 (99.7% of memory free)
Sorted by:

. describe, simple
```

```
gender    schtyp97  czship97  age99     hhinc99

. tabulate gender

    key!sex |
   (symbol) |
       1997 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,034       47.26       47.26
          2 |      1,154       52.74      100.00
------------+-----------------------------------
      Total |      2,188      100.00

. tabulate schtyp97

key!schoolt |
       ype2 |
   (symbol) |
       1997 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          2        0.09        0.09
          1 |         47        2.15        2.24
          2 |        376       17.18       19.42
          3 |      1,757       80.30       99.73
          4 |          5        0.23       99.95
          5 |          1        0.05      100.00
------------+-----------------------------------
      Total |      2,188      100.00
```

```
. tabulate schtyp97 gender, expected cchi2 row col cell all
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| chi2 contribution  |
|   row percentage   |
| column percentage  |
|  cell percentage   |
+--------------------+
key!school |
     type2 |
  (symbol) | key!sex (symbol) 1997
      1997 |         1          2 |     Total
-----------+----------------------+----------
         0 |         0          2 |         2
           |       0.9        1.1 |       2.0
           |       0.9        0.8 |       1.8
           |      0.00     100.00 |    100.00
           |      0.00       0.17 |      0.09
           |      0.00       0.09 |      0.09
-----------+----------------------+----------
         1 |        21         26 |        47
           |      22.2       24.8 |      47.0
           |       0.1        0.1 |       0.1
           |     44.68      55.32 |    100.00
           |      2.03       2.25 |      2.15
           |      0.96       1.19 |      2.15
-----------+----------------------+----------
         2 |       202        174 |       376
           |     177.7      198.3 |     376.0
           |       3.3        3.0 |       6.3
           |     53.72      46.28 |    100.00
           |     19.54      15.08 |     17.18
           |      9.23       7.95 |     17.18
-----------+----------------------+----------
         3 |       810        947 |     1,757
           |     830.3      926.7 |   1,757.0
           |       0.5        0.4 |       0.9
           |     46.10      53.90 |    100.00
           |     78.34      82.06 |     80.30
           |     37.02      43.28 |     80.30
-----------+----------------------+----------
         4 |         0          5 |         5
           |       2.4        2.6 |       5.0
           |       2.4        2.1 |       4.5
           |      0.00     100.00 |    100.00
           |      0.00       0.43 |      0.23
           |      0.00       0.23 |      0.23
-----------+----------------------+----------
         5 |         1          0 |         1
           |       0.5        0.5 |       1.0
           |       0.6        0.5 |       1.1
           |    100.00       0.00 |    100.00
           |      0.10       0.00 |      0.05
           |      0.05       0.00 |      0.05
-----------+----------------------+----------
```

```
     Total |     1,034      1,154 |     2,188
           |   1,034.0    1,154.0 |   2,188.0
           |       7.8        7.0 |      14.8
           |     47.26      52.74 |    100.00
           |    100.00     100.00 |    100.00
           |     47.26      52.74 |    100.00
          Pearson chi2(5) =  14.7625   Pr = 0.011
 likelihood-ratio chi2(5) =  17.8189   Pr = 0.003
               Cramér's V =   0.0821
                    gamma =   0.1268  ASE = 0.052
          Kendall's tau-b =   0.0514  ASE = 0.021
```

```
. generate school=schtyp97
```

*** CREATE A NEW VARIABLE CALLED SCHOOL**

```
. replace school=4 if (schtyp97==5)
(1 real change made)
```

*** COMBINE COLLEGE RESPONSES**

```
. replace school=. if (schtyp97==0)
(2 real changes made, 2 to missing)
```

*** SET NONRESPONSES TO MISSING**

*** CHECK CREATION OF SCHOOL VARIABLE**

```
. tabulate school schtyp97, missing
```

```
           |                  key!schooltype2 (symbol) 1997
    school |         0          1          2          3          4          5 |     Total
-----------+------------------------------------------------------------------+----------
         1 |         0         47          0          0          0          0 |        47
         2 |         0          0        376          0          0          0 |       376
         3 |         0          0          0      1,757          0          0 |     1,757
         4 |         0          0          0          0          5          1 |         6
         . |         2          0          0          0          0          0 |         2
-----------+------------------------------------------------------------------+----------
     Total |         2         47        376      1,757          5          1 |     2,188
```

```
. tabulate school gender, expected cchi2 row col cell all
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| chi2 contribution  |
|   row percentage   |
| column percentage  |
|  cell percentage   |
+--------------------+

           | key!sex (symbol) 1997
    school |         1          2 |     Total
-----------+----------------------+----------
         1 |        21         26 |        47
```

```
           |      22.2       24.8 |      47.0
           |       0.1        0.1 |       0.1
           |     44.68      55.32 |    100.00
           |      2.03       2.26 |      2.15
           |      0.96       1.19 |      2.15
-----------+----------------------+----------
         2 |       202        174 |       376
           |     177.9      198.1 |     376.0
           |       3.3        2.9 |       6.2
           |     53.72      46.28 |    100.00
           |     19.54      15.10 |     17.20
           |      9.24       7.96 |     17.20
-----------+----------------------+----------
         3 |       810        947 |     1,757
           |     831.1      925.9 |   1,757.0
           |       0.5        0.5 |       1.0
           |     46.10      53.90 |    100.00
           |     78.34      82.20 |     80.38
           |     37.05      43.32 |     80.38
-----------+----------------------+----------
         4 |         1          5 |         6
           |       2.8        3.2 |       6.0
           |       1.2        1.1 |       2.3
           |     16.67      83.33 |    100.00
           |      0.10       0.43 |      0.27
           |      0.05       0.23 |      0.27
-----------+----------------------+----------
     Total |     1,034      1,152 |     2,186
           |   1,034.0    1,152.0 |   2,186.0
           |       5.1        4.6 |       9.6
           |     47.30      52.70 |    100.00
           |    100.00     100.00 |    100.00
           |     47.30      52.70 |    100.00
Pearson chi2(3) =   9.6245   Pr = 0.022
 likelihood-ratio chi2(3) =   9.8515   Pr = 0.020
              Cramér's V =   0.0664
                   gamma =   0.1329  ASE = 0.052
          Kendall's tau-b =   0.0537  ASE = 0.021
```

```
. replace school=. if (school>=4)
(6 real changes made, 6 to missing)
```

*** SET COLLEGE TO MISSING -- INSUFFICIENT SAMPLE SIZE**

```
. tabulate school gender, expected cchi2 row col cell all
```

```
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| chi2 contribution  |
|   row percentage   |
| column percentage  |
|  cell percentage   |
+--------------------+

           | key!sex (symbol) 1997
    school |         1          2 |     Total
-----------+----------------------+----------
         1 |        21         26 |        47
           |      22.3       24.7 |      47.0
           |       0.1        0.1 |       0.1
           |     44.68      55.32 |    100.00
           |      2.03       2.27 |      2.16
           |      0.96       1.19 |      2.16
-----------+----------------------+----------
         2 |       202        174 |       376
           |     178.2      197.8 |     376.0
           |       3.2        2.9 |       6.1
           |     53.72      46.28 |    100.00
           |     19.55      15.17 |     17.25
           |      9.27       7.98 |     17.25
-----------+----------------------+----------
         3 |       810        947 |     1,757
           |     832.6      924.4 |   1,757.0
           |       0.6        0.6 |       1.2
           |     46.10      53.90 |    100.00
           |     78.41      82.56 |     80.60
           |     37.16      43.44 |     80.60
-----------+----------------------+----------
     Total |     1,033      1,147 |     2,180
           |   1,033.0    1,147.0 |   2,180.0
           |       3.9        3.5 |       7.4
           |     47.39      52.61 |    100.00
           |    100.00     100.00 |    100.00
           |     47.39      52.61 |    100.00

          Pearson chi2(2) =   7.3581   Pr = 0.025
 likelihood-ratio chi2(2) =   7.3490   Pr = 0.025
               Cramér's V =   0.0581
                    gamma =   0.1245  ASE = 0.052
          Kendall's tau-b =   0.0501  ASE = 0.021

.
end of do-file
```