# CHAPTER 25: Two Categorical Variables: The Chi-Square Test

**Basic Practice of Statistics**
7th Edition

**Lecture PowerPoint Slides**

# In Chapter 25, We Cover …

- Two-way tables
- The problem of multiple comparisons
- Expected counts in two-way tables
- The chi-square test statistic
- Cell counts required for the chi-square test
- Using technology
- Uses of the chi-square test: independence and homogeneity
- The chi-square distributions
- The chi-square test for goodness of fit*

# Two-Way Tables

- The two-sample *z* procedures of Chapter 18 allow us to compare the proportions of successes in two groups, either two populations or two treatment groups in an experiment.
- When there are more than two outcomes, or when we want to compare more than two groups, we need a new statistical test.
- The new test addresses a general question: Is there a relationship between two categorical variables?
- Two-way tables of counts can be used to describe relationships between any two categorical variables.

# Two-Way Tables—Example

A sample survey asked a random sample of young adults, “Where do you live now?”  The table below is a two-way table of all 2984 people in the sample (both men and women) classified by their age and by where they live.  Living arrangement is a categorical variable. Even though age is quantitative, the two-way table treats age as dividing young adults into four categories.  Here is a table that summarizes the data:

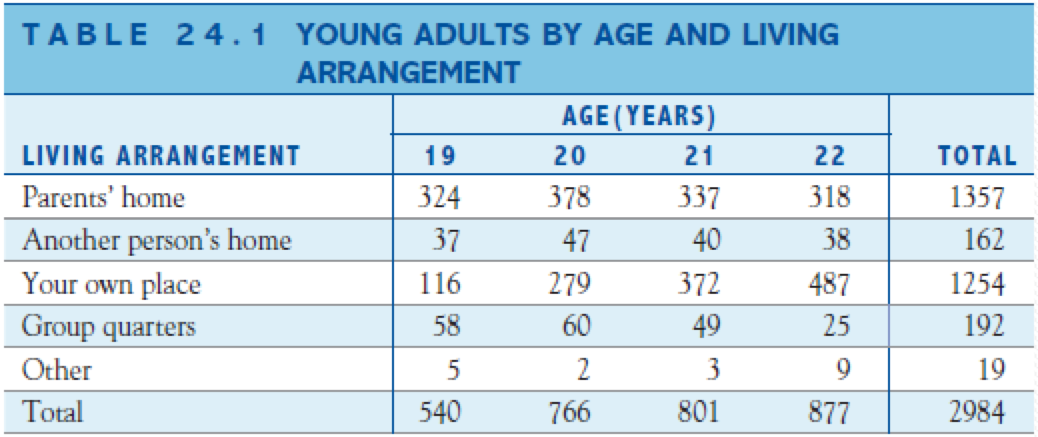


TABLE 24.1 YOUNG ADULTS BY AGE AND LIVING ARRANGEMENT

| | AGE (YEARS) | | | | |
|---|---|---|---|---|---|
| LIVING ARRANGEMENT | 19 | 20 | 21 | 22 | TOTAL |
| Parents’ home | 324 | 378 | 337 | 318 | 1357 |
| Another person’s home | 37 | 47 | 40 | 38 | 162 |
| Your own place | 116 | 279 | 372 | 487 | 1254 |
| Group quarters | 58 | 60 | 49 | 25 | 192 |
| Other | 5 | 2 | 3 | 9 | 19 |
| Total | 540 | 766 | 801 | 877 | 2984 |

**PROBLEM:**

(a) Calculate the conditional distribution (in proportions) of the living arrangement for each age.

(b) Make an appropriate graph for comparing the conditional distributions in part (a).

(c) Are the distributions of living arrangements under the four ages similar or different? Give appropriate evidence from parts (a) and (b) to support your answer.

# Two-Way Tables—Example

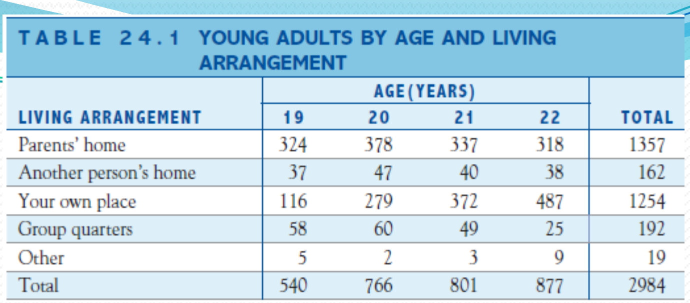


TABLE 24.1 YOUNG ADULTS BY AGE AND LIVING ARRANGEMENT

| | AGE (YEARS) | | | | |
|---|---|---|---|---|---|
| LIVING ARRANGEMENT | 19 | 20 | 21 | 22 | TOTAL |
| Parents' home | 324 | 378 | 337 | 318 | 1357 |
| Another person's home | 37 | 47 | 40 | 38 | 162 |
| Your own place | 116 | 279 | 372 | 487 | 1254 |
| Group quarters | 58 | 60 | 49 | 25 | 192 |
| Other | 5 | 2 | 3 | 9 | 19 |
| Total | 540 | 766 | 801 | 877 | 2984 |

For 19-year olds, the distribution of living arrangements was:

Parents' home, $\frac{324}{540} = 60.0\%$; Another person's home, $\frac{37}{540} = 6.9\%$; Your own Place, $\frac{116}{540} = 21.5\%$; Group quarters, $\frac{58}{540} = 10.7\%$; and Other, $\frac{5}{540} = 0.9\%$

For 20-year olds, the distribution was:

Parents' home, $\frac{378}{766} = 49.3\%$; Another person's home, $\frac{47}{766} = 6.1\%$; Your own Place, $\frac{279}{766} = 36.4\%$; Group quarters, $\frac{60}{766} = 7.8\%$; and Other, $\frac{2}{766} = 0.3\%$

For 21-year olds, the distribution was:

Parents' home, $\frac{337}{801} = 42.1\%$; Another person's home, $\frac{40}{801} = 5.0\%$; Your own Place, $\frac{372}{801} = 46.4\%$; Group quarters, $\frac{49}{801} = 6.1\%$; and Other, $\frac{3}{801} = 0.4\%$

And for 22-year olds, the distribution was:

Parents' home, $\frac{318}{877} = 36.3\%$; Another person's home, $\frac{38}{877} = 4.3\%$; Your own Place, $\frac{487}{877} = 55.5\%$; Group quarters, $\frac{25}{877} = 2.9\%$; and Other, $\frac{9}{877} = 1.0\%$

# Two-Way Tables—Example (cont’d)

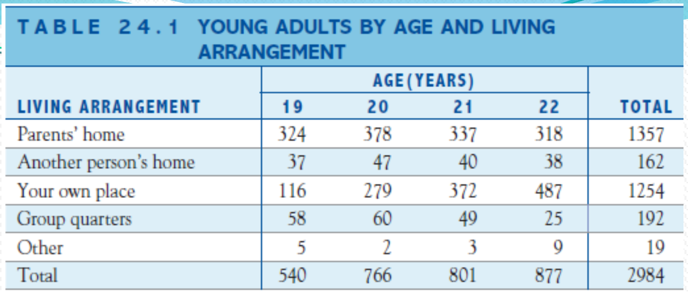
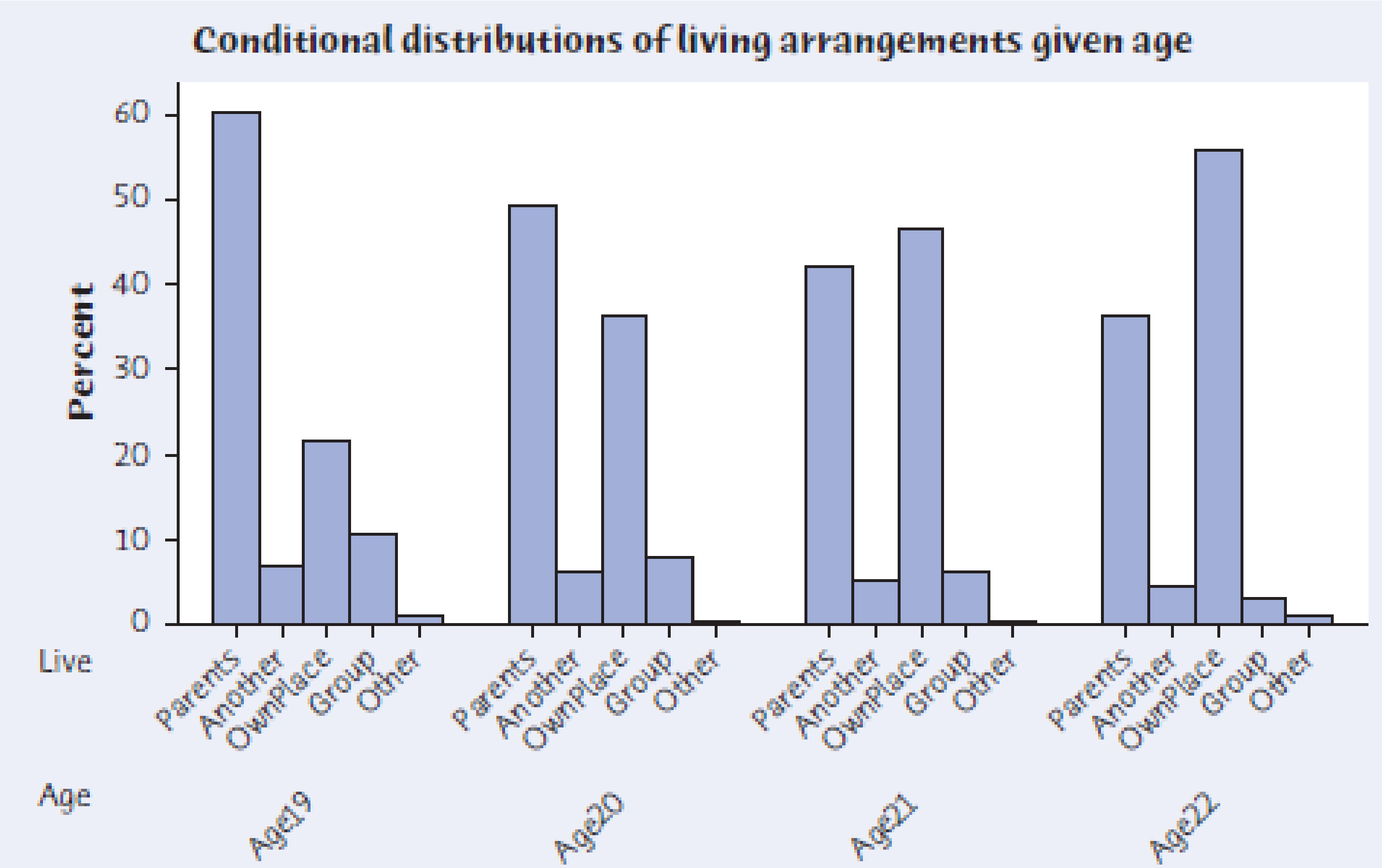


**TABLE 24.1 YOUNG ADULTS BY AGE AND LIVING ARRANGEMENT**

| | AGE (YEARS) | | | | |
|---|---|---|---|---|---|
| LIVING ARRANGEMENT | 19 | 20 | 21 | 22 | TOTAL |
| Parents’ home | 324 | 378 | 337 | 318 | 1357 |
| Another person’s home | 37 | 47 | 40 | 38 | 162 |
| Your own place | 116 | 279 | 372 | 487 | 1254 |
| Group quarters | 58 | 60 | 49 | 25 | 192 |
| Other | 5 | 2 | 3 | 9 | 19 |
| Total | 540 | 766 | 801 | 877 | 2984 |



Conditional distributions of living arrangements given age

Percent within levels of Age.

# The Problem of Multiple Comparisons

- To address the general question of whether there is a relationship between two categorical variables, we look for significant differences among the conditional distributions of one categorical variable given the values of the other variable.
- The null hypothesis is that there is no relationship between two categorical variables:

  $H_0$: there is no difference in the distribution of a categorical variable for several populations or treatments.
- The alternative hypothesis says that there is a relationship, but it does not specify any particular kind of relationship:

  $H_\mathrm{a}$:  there is a difference in the distribution of a categorical variable for several populations or treatments.
- We could compare many pairs of proportions, ending up with many tests and many *P*-values—***BAD IDEA!***
- *When we do many individual tests or confidence intervals, the individual P-values and confidence levels don't tell us how confident we can be in all of the inferences taken together.*

# The Problem of Multiple Comparisons

- The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions is common in statistics. This is the problem of multiple comparisons. Statistical methods for dealing with multiple comparisons usually have two parts:
  1. An *overall test* to see if there is good evidence of any differences among the parameters that we want to compare
  2. A detailed *follow-up analysis* to decide which of the parameters differ and to estimate how large the differences are
- The overall test, though more complex than the tests we met earlier, is reasonably straightforward. The follow-up analysis can be quite elaborate.

# Expected Counts in Two-Way Tables

- Our general null hypothesis $H_0$ is that there is no relationship between the two categorical variables that label the rows and columns of a two-way table.
- To test $H_0$, we compare the observed counts in the table with the expected counts, the counts we would expect (except for random variation) if $H_0$ were true.
- If the observed counts are far from the expected counts, that is evidence against $H_0$.

**EXPECTED COUNTS**

- The expected count in any cell of a two-way table when $H_0$ is true is

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

# Expected Counts in Two-Way Tables —Example

Finding the expected counts is not that difficult, as the following example illustrates.

The null hypothesis in the age and living arrangements study is that there is no difference in the distribution of living arrangements, whether it’s a 19-, 20-, 21-, or 22-year-old.

To find the expected counts, we start by assuming that $H_0$ is true. We can see from the two-way table that 1357 of the 2984 young adults surveyed lived in their parents’ homes.

If the age of the young adult has *no effect* on their chosen living arrangement, the proportion of those living at their parents’ home for each age should be 1357/2984 = 0.455.

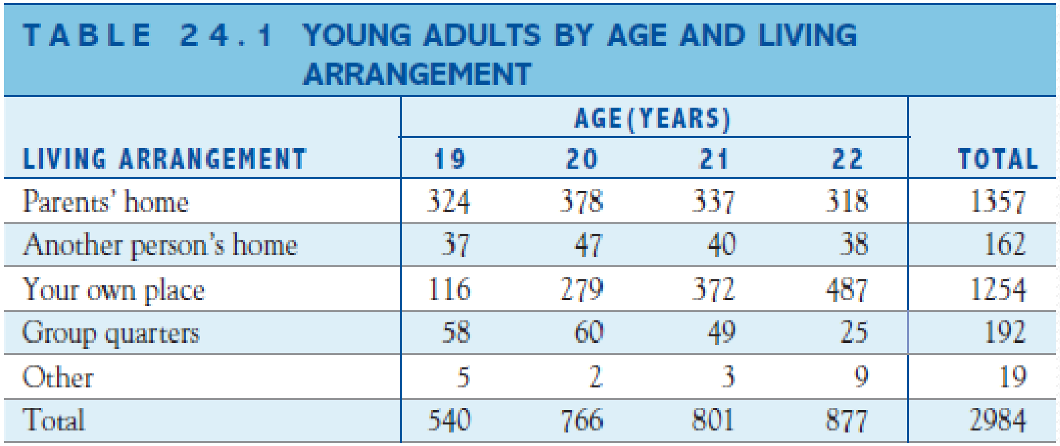


TABLE 24.1 YOUNG ADULTS BY AGE AND LIVING ARRANGEMENT

| | AGE (YEARS) | | | | |
|---|---|---|---|---|---|
| LIVING ARRANGEMENT | 19 | 20 | 21 | 22 | TOTAL |
| Parents' home | 324 | 378 | 337 | 318 | 1357 |
| Another person's home | 37 | 47 | 40 | 38 | 162 |
| Your own place | 116 | 279 | 372 | 487 | 1254 |
| Group quarters | 58 | 60 | 49 | 25 | 192 |
| Other | 5 | 2 | 3 | 9 | 19 |
| Total | 540 | 766 | 801 | 877 | 2984 |

# Expected Counts in Two-Way Tables—Example (cont'd)

The overall proportion of young adults living in their parents' homes was 1357/2984 = 0.455. So the expected counts of those living at their parents' homes for each age are: 19-yr olds, $540\left(\frac{1357}{2984}\right) = 245.57$; 20-year olds, $766\left(\frac{1357}{2984}\right) = 348.35$; etc.

The overall proportion of young adults living in another person's home was 162/2984 = 0.0543. So the expected counts of those living in another person's home for each age are: 19-yr olds, $540\left(\frac{162}{2984}\right) = 29.32$; 20-year olds, $766\left(\frac{162}{2984}\right) =$ $41.59$; etc.

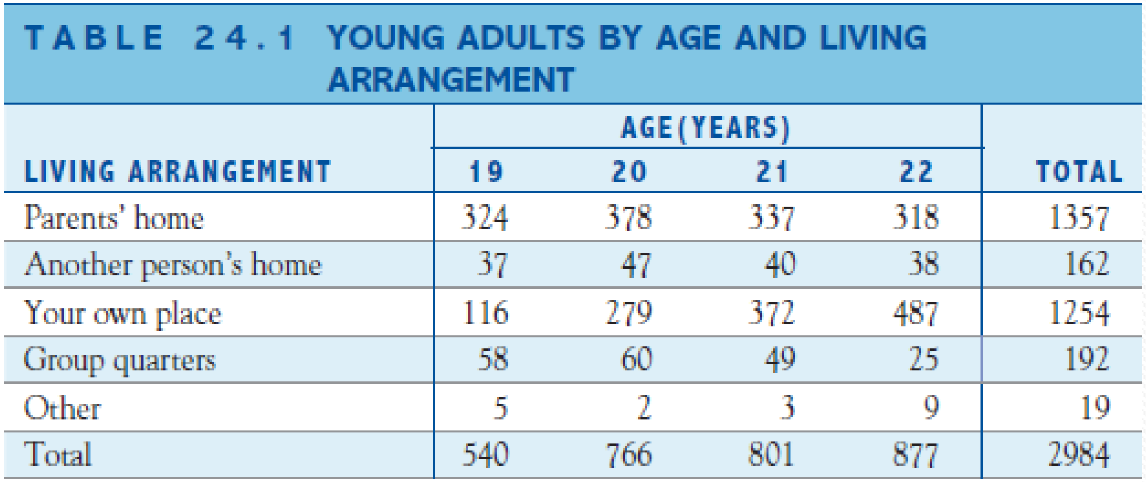


**TABLE 24.1 YOUNG ADULTS BY AGE AND LIVING ARRANGEMENT**

| | AGE (YEARS) | | | | |
|---|---|---|---|---|---|
| **LIVING ARRANGEMENT** | **19** | **20** | **21** | **22** | **TOTAL** |
| Parents' home | 324 | 378 | 337 | 318 | 1357 |
| Another person's home | 37 | 47 | 40 | 38 | 162 |
| Your own place | 116 | 279 | 372 | 487 | 1254 |
| Group quarters | 58 | 60 | 49 | 25 | 192 |
| Other | 5 | 2 | 3 | 9 | 19 |
| Total | 540 | 766 | 801 | 877 | 2984 |

# The Chi-Square Statistic

- To test whether the observed differences among the conditional distributions are statistically significant, we compare the observed and expected counts. The test statistic that makes the comparison is the *chi-square statistic*.

**CHI-SQUARE STATISTIC**

- The chi-square statistic is a measure of how far the observed counts in a two-way table are from the expected counts if $H_0$ were true. The formula for the statistic is

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

- The sum is over all cells in the table.

# Cell Counts Required for the Chi-Square Test

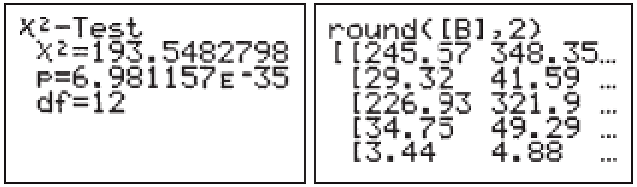
- The chi-square test, like the *z* procedures for comparing two proportions, is an approximate method that becomes more accurate as the counts in the cells of the table get larger.
- Fortunately, the chi-square approximation is accurate for quite modest counts.

**CELL COUNTS REQUIRED FOR THE CHI-SQUARE TEST**

- You can safely use the chi-square test with critical values from the chi-square distribution when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater. In particular, all four expected counts in a 2 × 2 table should be 5 or greater.
- Note that the guideline uses *expected* cell counts.

# Using Technology

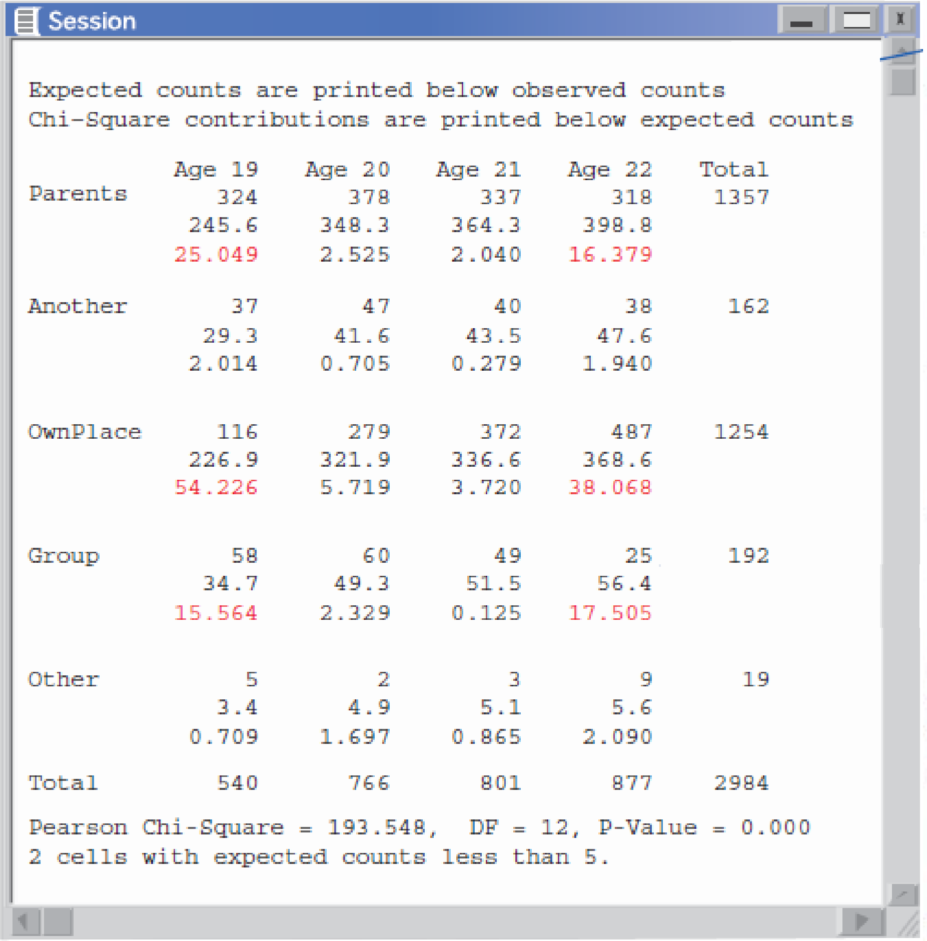
## Texas Instruments Graphing Calculator



```
χ²-Test
 χ²=193.5482798
 p=6.981157ᴇ-35
 df=12
```

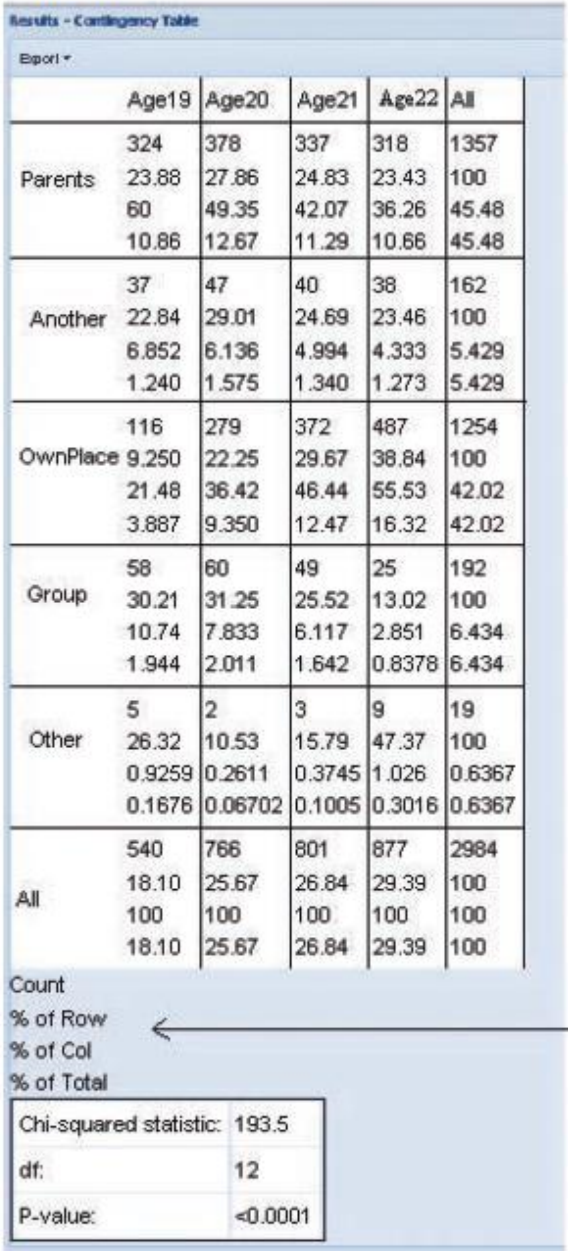
```
round([B],2)
[[245.57  348.35…
 [29.32   41.59  …
 [226.93  321.9  …
 [34.75   49.29  …
 [3.44    4.88   …
```

## Minitab



```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

          Age 19   Age 20   Age 21   Age 22   Total
Parents      324      378      337      318    1357
           245.6    348.3    364.3    398.8
          25.049    2.525    2.040   16.379

Another       37       47       40       38     162
            29.3     41.6     43.5     47.6
           2.014    0.705    0.279    1.940

OwnPlace     116      279      372      487    1254
           226.9    321.9    336.6    368.6
          54.226    5.719    3.720   38.068

Group         58       60       49       25     192
            34.7     49.3     51.5     56.4
          15.564    2.329    0.125   17.505

Other          5        2        3        9      19
             3.4      4.9      5.1      5.6
           0.709    1.697    0.865    2.090

Total        540      766      801      877    2984

Pearson Chi-Square = 193.548,  DF = 12, P-Value = 0.000
2 cells with expected counts less than 5.
```

## CrunchIt!



| | Age19 | Age20 | Age21 | Age22 | All |
|---|---|---|---|---|---|
| Parents | 324<br>23.88<br>60<br>10.86 | 378<br>27.86<br>49.35<br>12.67 | 337<br>24.83<br>42.07<br>11.29 | 318<br>23.43<br>36.26<br>10.66 | 1357<br>100<br>45.48<br>45.48 |
| Another | 37<br>22.84<br>6.852<br>1.240 | 47<br>29.01<br>6.136<br>1.575 | 40<br>24.69<br>4.994<br>1.340 | 38<br>23.46<br>4.333<br>1.273 | 162<br>100<br>5.429<br>5.429 |
| OwnPlace | 116<br>9.250<br>21.48<br>3.887 | 279<br>22.25<br>36.42<br>9.350 | 372<br>29.67<br>46.44<br>12.47 | 487<br>38.84<br>55.53<br>16.32 | 1254<br>100<br>42.02<br>42.02 |
| Group | 58<br>30.21<br>10.74<br>1.944 | 60<br>31.25<br>7.833<br>2.011 | 49<br>25.52<br>6.117<br>1.642 | 25<br>13.02<br>2.851<br>0.8378 | 192<br>100<br>6.434<br>6.434 |
| Other | 5<br>26.32<br>0.9259<br>0.1676 | 2<br>10.53<br>0.2611<br>0.06702 | 3<br>15.79<br>0.3745<br>0.1005 | 9<br>47.37<br>1.026<br>0.3016 | 19<br>100<br>0.6367<br>0.6367 |
| All | 540<br>18.10<br>100<br>18.10 | 766<br>25.67<br>100<br>25.67 | 801<br>26.84<br>100<br>26.84 | 877<br>29.39<br>100<br>29.39 | 2984<br>100<br>100<br>100 |

Count
% of Row
% of Col
% of Total

| | |
|---|---|
| Chi-squared statistic: | 193.5 |
| df: | 12 |
| P-value: | <0.0001 |

This key identifies the output for each cell in the table

# Using Technology

- The chi-square test is an overall test for detecting relationships between two categorical variables. If the test is significant, it is important to look at the data to learn the nature of the relationship. We have three ways to look at the data:

1. Compare selected percents: Which cells occur in quite different percents in the different conditional distributions?
2. Compare observed and expected cell counts: Which cells have more or fewer observations than we would expect if $H_0$ were true?
3. Look at the terms of the chi-square statistic: Which cells contribute the most to the value of $\chi^2$?

# Uses of the Chi-Square Test: Independence and Homogeneity

- The test we have been using to this point is generally referred to as the **chi-square test for independence**, as thus far all the examples have been questions about whether two classification variables are independent or not.
- In a different setting for a two-way table, in which we compare separate samples from two or more populations, or from two or more treatments in a randomized controlled experiment, “which population" is now one of the variables for the two-way table.
- For each sample, we classify individuals according to one variable, and we are interested in whether or not the probabilities of being classified in each category of this variable are the same for each population.
- In this context, our calculations for the chi-square test are unchanged, but the method of collecting the data is different.
- This use of the chi-square test is referred to as the **chi-square test for homogeneity** since we are interested in whether or not the populations from which the samples are selected are homogeneous (the same) with respect to the single classification variable.

# Uses of the Chi-Square Test: Independence and Homogeneity

**USES OF THE CHI-SQUARE TEST**

- Use the chi-square test to test the null hypothesis

  $H_0$ : there is no relationship between two categorical variables

  when you have a two-way table from one of these situations:

  - A single SRS, with each individual classified according to both of two categorical variables. In this case, the null hypothesis of no relationship says that the two categorical variables are independent and the test is called the **chi-square test of independence**.
  - Independent SRSs from two or more populations, with each individual classified according to one categorical variable. (The other variable says which sample the individual comes from.) In this case, the null hypothesis of no relationship says the populations are homogeneous and the test is called the **chi-square test of homogeneity**.

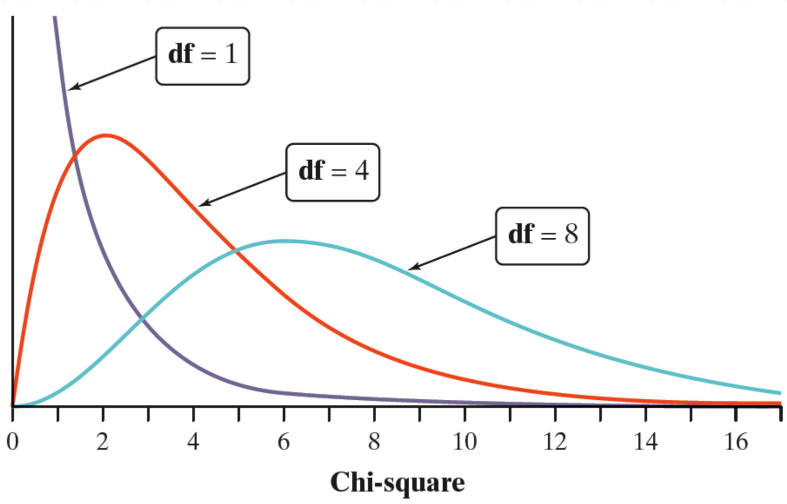
# The Chi-Square Distributions

- Software usually finds *P*-values for us. The *P*-value for a chi-square test comes from comparing the value of the chi-square statistic with critical values for a chi-square distribution.

**THE CHI-SQUARE DISTRIBUTIONS**

- The chi-square distributions are a family of distributions that take only positive values and are skewed to the right. A specific chi-square distribution is specified by giving its degrees of freedom.

- The chi-square test for a two-way table with *r* rows and *c* columns uses critical values from the chi-square distribution with (*r* – 1)(*c* – 1) degrees of freedom. The *P*-value is the area under the density curve of this chi-square distribution to the right of the value of test statistic.


df = 1
df = 4
df = 8
Chi-square