# Chapter 22. Two Categorical Variables: The Chi-Square Test

**Topics covered in this chapter:**

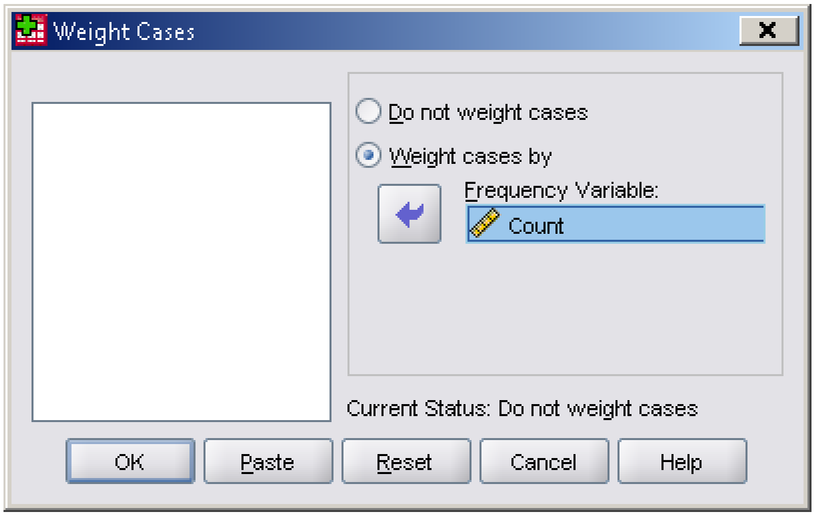
- **Two-Way Tables**
- **The Chi-Square Test**

## Two-Way Tables

**Example 22.1: Where do young people live?**

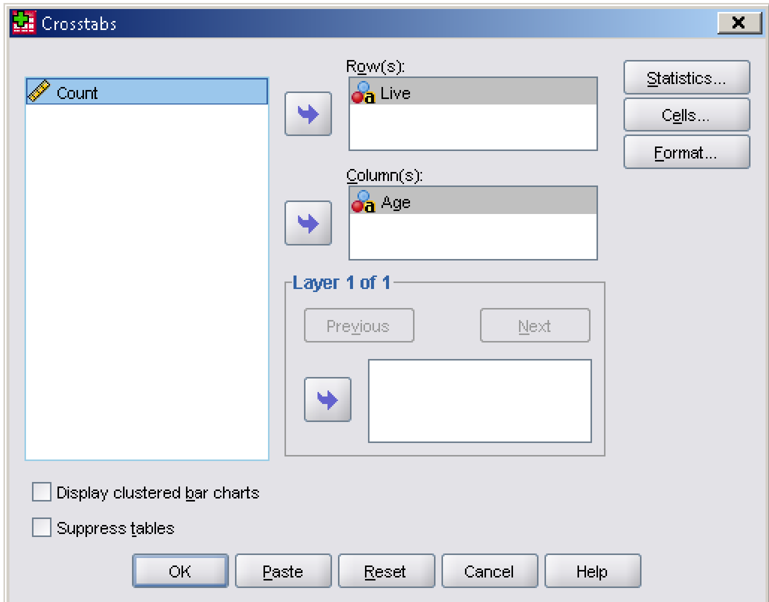
**The Problem:** A sample survey asked a random sample of young adults, “Where do you live now?” How does living arrangement vary by the age of the young person? Even though age is quantitative, the two-way table treats age as a categorical variables by dividing the young people into four age groups.

1. Open the data set *ta22-01.por*. Notice that all combinations of ***Live*** and ***Age*** are listed in the first two columns. The count of young people is given in the third column.
2. Click **Data** then **Weight Cases**.
3. Click on **Weight cases by** and move ***Count*** into the **Frequency Variable** box. Click **OK**.

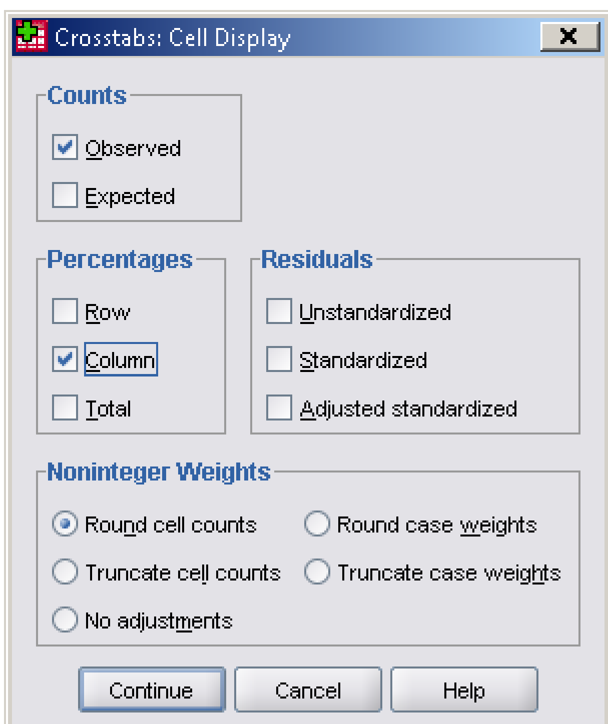



4. Change counts to percents.
   a. Click **Analyze**, then **Descriptive Statistics**, then **Crosstabs**.

b. Move ***Live*** into the **Row(s)** box.
c. Move ***Age*** into the **Column(s)** box.




d. Click the **Cells** button.
e. Under **Percentages** put a check next to **Column**.



f. Click **Continue.**
g. Click **OK**. A new window will pop up with your output.

**Live * Age Crosstabulation**

| | | | Age | | | | |
|---|---|---|---|---|---|---|---|
| | | | Age19 | Age20 | Age21 | Age22 | Total |
| Live | Another | Count | 37 | 47 | 40 | 38 | 162 |
| | | % within Age | 6.9% | 6.1% | 5.0% | 4.3% | 5.4% |
| | Group | Count | 58 | 60 | 49 | 25 | 192 |
| | | % within Age | 10.7% | 7.8% | 6.1% | 2.9% | 6.4% |
| | Other | Count | 5 | 2 | 3 | 9 | 19 |
| | | % within Age | .9% | .3% | .4% | 1.0% | .6% |
| | OwnPlac | Count | 116 | 279 | 372 | 487 | 1254 |
| | | % within Age | 21.5% | 36.4% | 46.4% | 55.5% | 42.0% |
| | Parents | Count | 324 | 378 | 337 | 318 | 1357 |
| | | % within Age | 60.0% | 49.3% | 42.1% | 36.3% | 45.5% |
| | Total | Count | 540 | 766 | 801 | 877 | 2984 |
| | | % within Age | 100.0% | 100.0% | 100.0% | 100.0% | 100.0% |

## The Chi-Square Test

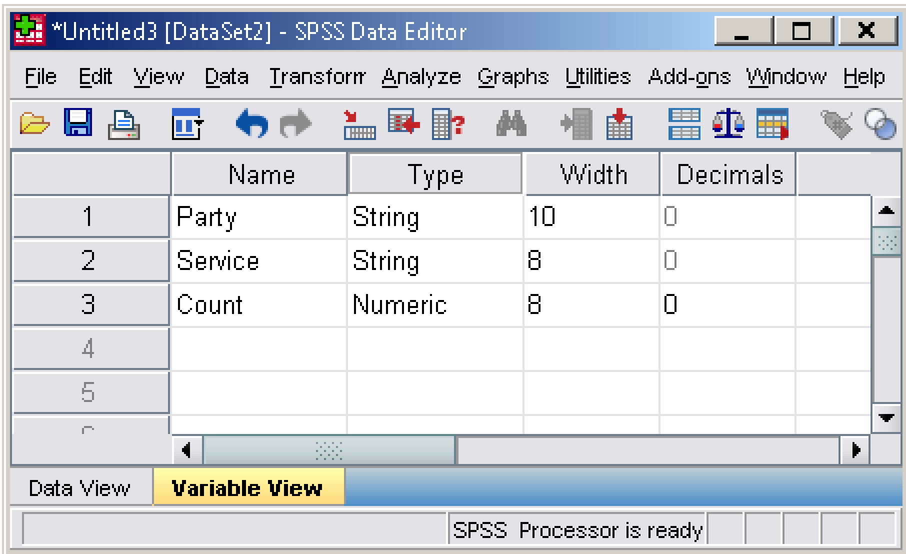
**Example 22.6: Are cell-only telephone users different?**

**The Problem:** Random digit dialing telephone surveys do not call cell phone numbers. If the opinions of people who have only cell phones differ from those of people who still have the landline service, the poll results may not represent the entire adult populations. The Pew Research Center interviewed separate random samples of cell-only and landline telephone users. In SPSS carry out a chi-square test for:

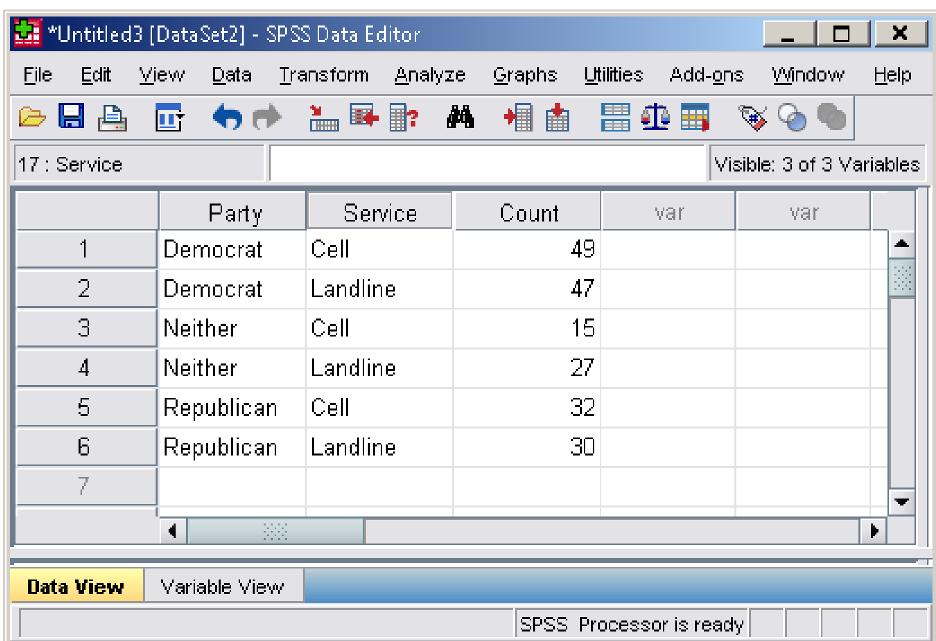
Ho: There is no relationship between type of phone service and political party affiliation.
Ha: There is some relationship type of phone service and political party affiliation.

1. Enter data into SPSS.
   a. Go to **Variable View**.
   b. Under **Name** in row 1, type ***Party***, corresponding to the political party affiliation. Under **Type**, select **String**. Under Width, select “10”.
   c. Under **Name** in row 2, type ***Service***, corresponding to the type of phone service. Under **Type**, select **String**.

d. Under **Name** in row 3, type ***Count***, corresponding to the number of phone users that correspond to the service type and political party affiliation of that row. Under **Type**, select **Numeric**.

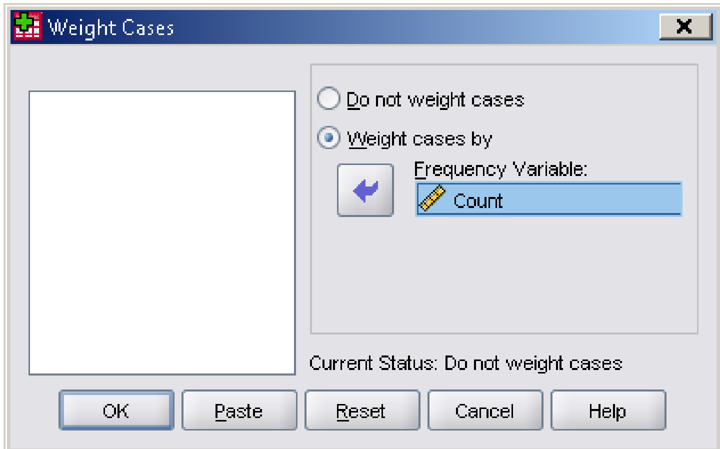



e. Go to **Data View** and enter the data.

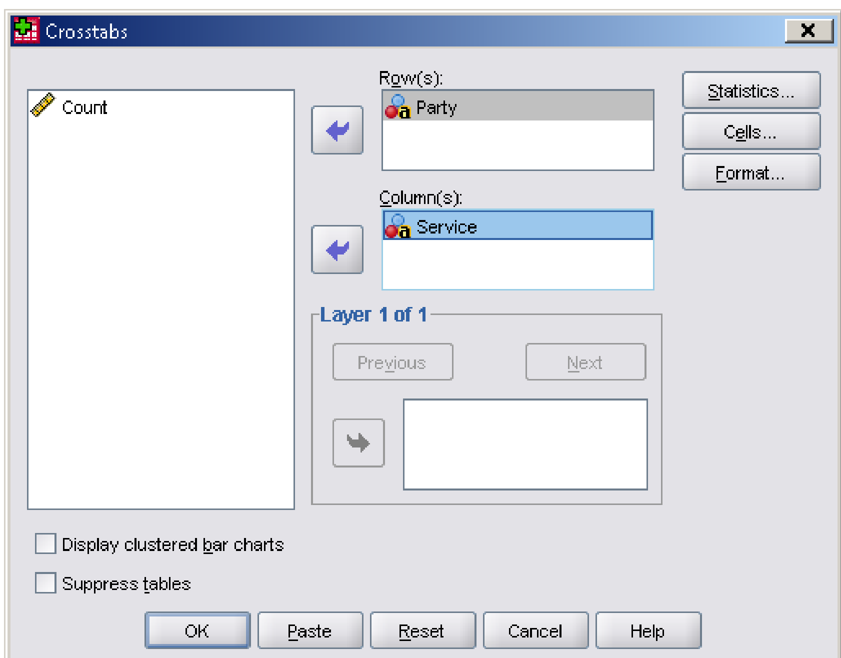



g. Click **Data** then **Weight Cases**.
h. Click on **Weight cases by** and move ***Count*** into the **Frequency Variable** box.

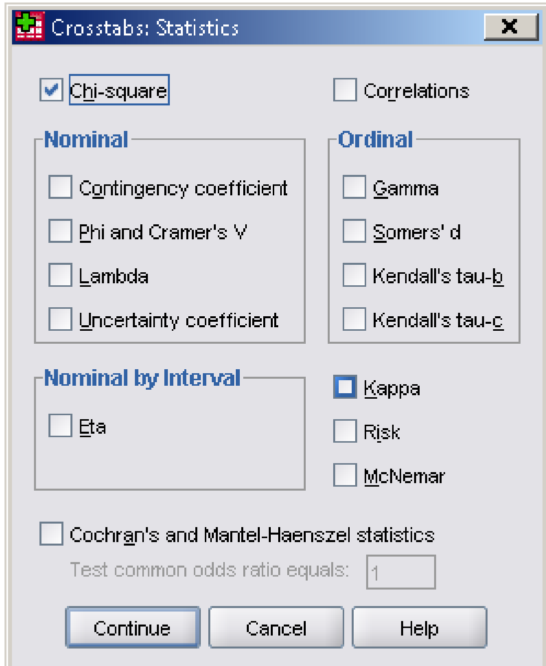
i. Click **OK**.



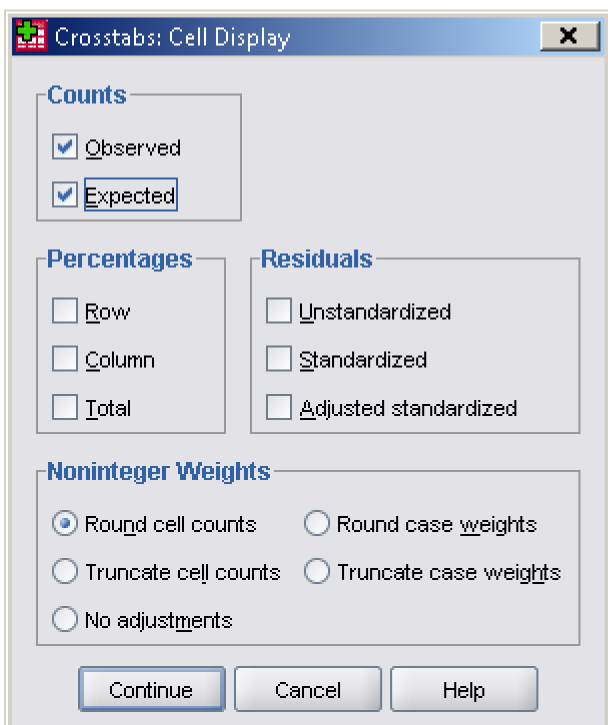

2. Perform the chi-square test.
   a. Click **Analyze,** scroll down to **Descriptive Statistics**, then click on **Crosstabs**.
   b. Move ***Party*** into the **Row(s)** box.
   c. Move ***Service*** into the **Column(s)** box.




   d. Click the **Statistics** button at the right side of the window.
   e. Put a check in the box in front of **Chi-square**.
   f. Click **Continue**.




g. To include expected cell counts in your **Crosstabulation** table in your output, click **Cells**, and under **Counts** put a check mark next to **Expected.**



h. Click **Continue.** Then click **OK**.
i. A new window will pop up with your output.

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| Grade * Activities | 119 | 100.0% | 0 | .0% | 119 | 100.0% |

**Grade * Activities Crosstabulation**

| | | | Activities | | | |
|---|---|---|---|---|---|---|
| | | | 2 to 12 hours | Less than 2 hours | More than 12 hours | Total |
| Grade | C or better | Count | 68 | 11 | 3 | 82 |
| | | Expected Count | 62.7 | 13.8 | 5.5 | 82.0 |
| | D or F | Count | 23 | 9 | 5 | 37 |
| | | Expected Count | 28.3 | 6.2 | 2.5 | 37.0 |
| Total | | Count | 91 | 20 | 8 | 119 |
| | | Expected Count | 91.0 | 20.0 | 8.0 | 119.0 |

**Chi-Square Tests**

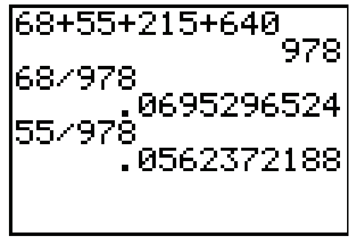
| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 6.926[a] | 2 | .031 |
| Likelihood Ratio | 6.520 | 2 | .038 |
| N of Valid Cases | 119 | | |

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 2.49.

## Chapter 22 Exercises

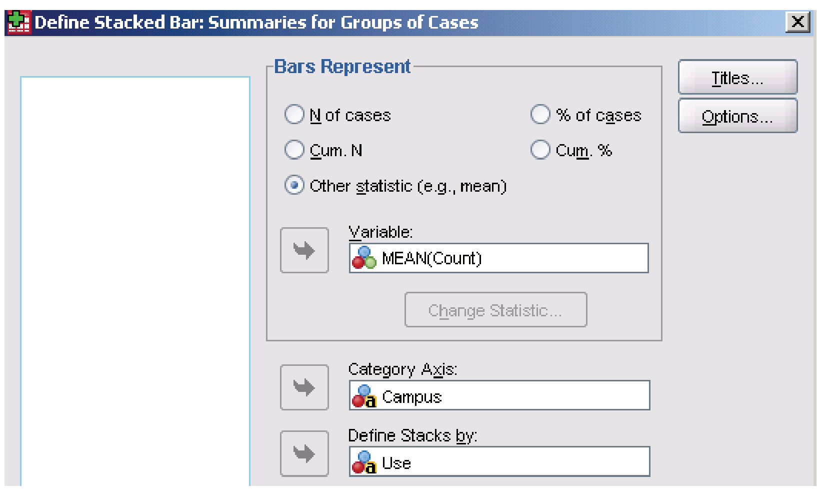
22.1 Facebook at Penn State.
22.3 Attitudes towards recycled products.
22.5 Facebook at Penn State.
22.13 Saving birds from windows.
22.15 Police harassment?
22.17 What's your sign?
22.29 Free speech for racists?
22.43 How are schools doing?
22.45 Market research.
22.47 Party support in brief.

**Chapter 22 SPSS Solutions**

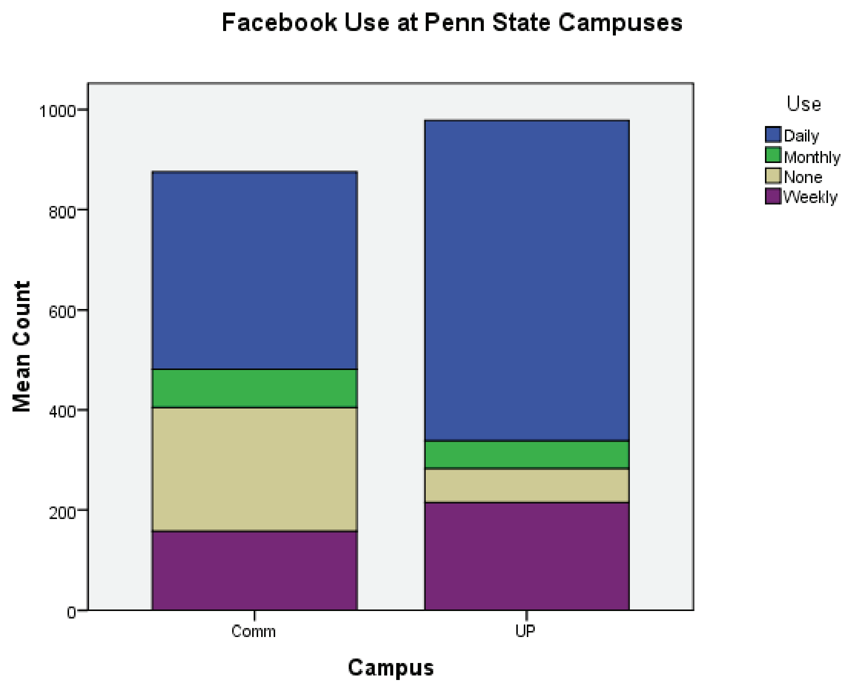



**22.1** To find the percent of University Park students who fall in each Facebook category, add the values given for University Park (68 + 55 + 215 + 640 = 978). Then, divide each category's number by the total. We see that about 7% of the University Park students do not use Facebook and about 5.6% use it several times per month or less. Continue with the other two categories, to find that about 22% use use Facebook at least once a week and 65.4% use it at least once a day.

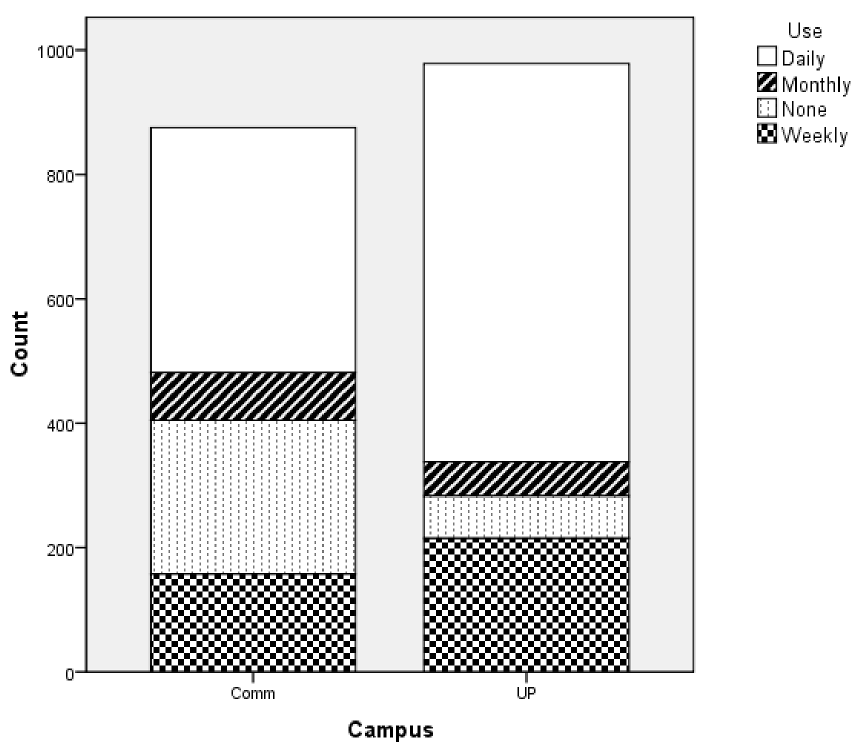
To compare the distributions, we'll make a stacked bar chart of the data with one bar for each of the University Park and Commonwealth students. Open data file *ex22-01*. To create the chart, click **Graphs**, **Legacy Dialogs**, **Bar**. We want the **Stacked** bar chart where data are **Summaries for groups of cases**. We'll create an initial chart (and modify it later with the Chart Editor) by defining it as below (don't forget to give your graph a **Titles**).




It's hard to compare the two distributions in the initial graph, because there were different numbers of students surveyed at the different campuses.

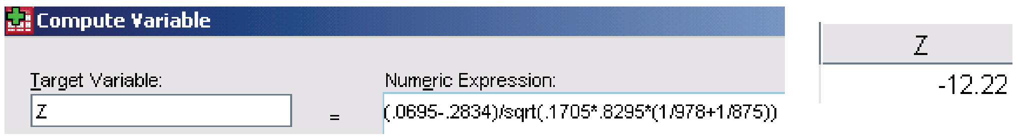



Click in the graph to bring up the Chart Editor, then click **Options**, **Scale to 100%**. You can also click in the *y*-axis label and remove it by unchecking the **Display axis title** box (with percents showing, this is not needed). If you wish, click the **Variables** tab and change the **Style** for **Use** from Color to pattern. **Apply** and **Close** the Chart Editor.



It is clear that University Park students are much more likely to be daily Facebook users; Commonwealth students are more likely to not use it at all; the "occasional" users seem similar.

**22.3** Parts (a) and (b) want us to compute tests for a difference in proportions. We first compute the test for those who do not use Facebook. There were 68/978 = 0.0695 University Park students who do not use it and 248/875 = 0.2834 Commonwealth students who do not. The pooled proportion is (68+248)/(978+875) = 0.1705.




With a test statistic of $z = -12.22$, we do not really need to compute the $P$-value, as this will be (essentially) 0. There is a difference. University Park students are definitely more likely to use Facebook.

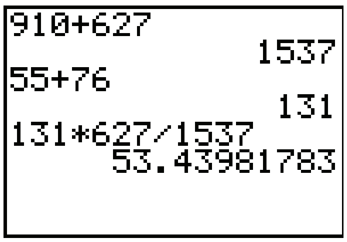
We repeat the computation for those using Facebook at least once a week. The observed proportions are: University Park, 215/978 = 0.2198 and Commonwealth, 157/875 = 0.1794. The pooled proportion is (215+157)/(978+875) = 0.2008.

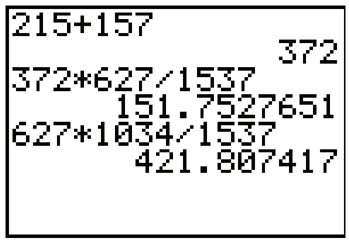


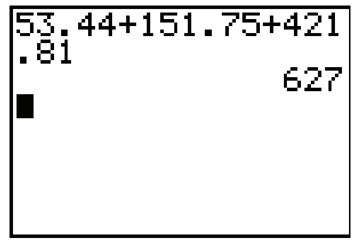
The difference is not quite as significant, but is still significant at the 0.05 level ($P$-value 0.030). Again, University Park students are more likely to use Facebook at least once a week.

These two $P$-values can't tell use about the two distributions for all four outcomes because they don't represent all the categories. Further, they are really dependent – if a student is in one category, they can't be in another, but we don't know which other category.

**22.5** If there is no relationship, the expected counts are (R x C)/T, where R is the row total, C is the column total, and T is the grand total. The grand total for the table is 910 + 627 = 1537. There were a total of 55 + 76 = 131 students who use Facebook several times a month or less. The expected count of these for Commonwealth students is 53.44. Similarly, the Commonwealth expected count for at least weekly users is 151.75 and for at least once a day users, the expected count is 421.81. The expected counts should total 627; we see they do.

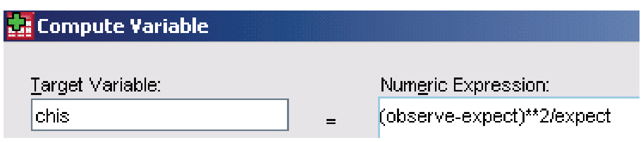


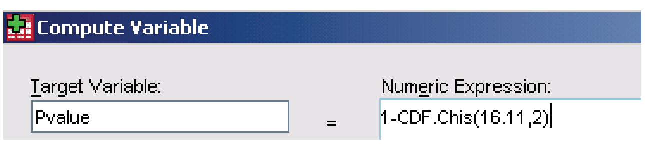




The general trend for these older Commonwealth students is that they are more likely to be occasional Facebook users than daily users; other claims on their time is most likely the reason.

**22.13** The expected counts are 53*1/3 = 17.6667, since if the tilts made no difference, there should be an equal number of strikes on each type of window. We enter the observed and expected counts in two variables and compute the components of the chi-square statistic as shown below. Sum the components to find $\chi^2 = 16.11$.



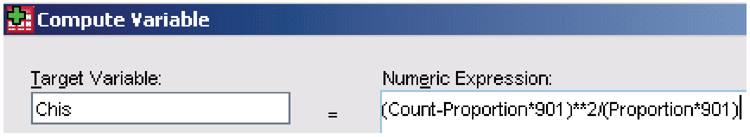

| observe | expect | chis |
|---|---|---|
| 31 | 17.6667 | 10.06 |
| 14 | 17.6667 | 0.76 |
| 8 | 17.6667 | 5.29 |



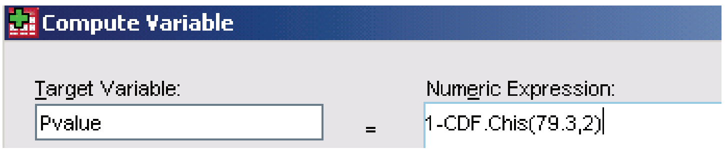

| Pvalue |
|---|
| 0.0003 |

**22.15** We entered the data as shown at right. Our null hypothesis is that the counts agree with the population proportions; the alternate is that they do not agree. SPSS still doesn't like summarized data. We add the number of observations to find that 401 + 480 + 20 = 901 citations represented. We compute the test statistic entries (and then sum them) to find $\chi^2 = 79.3$.

| Proportion | Count | Age |
|---|---|---|
| 0.328 | 401 | 16 to 29 |
| 0.594 | 480 | 30 to 59 |
| 0.078 | 20 | 60 up |




| Chis |
|---|
| 37.64 |
| 5.69 |
| 35.97 |

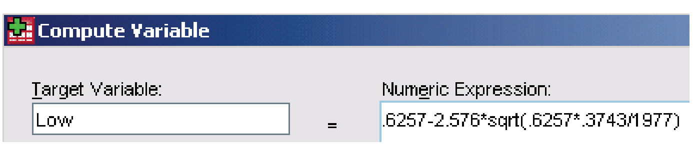



| Pvalue |
|---|
| 0.0000 |

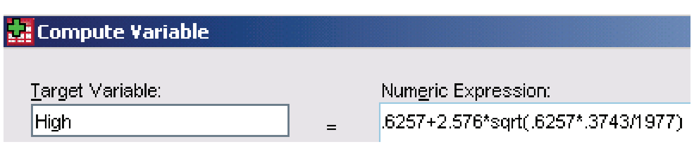
With a test statistic of $\chi^2 = 119.84$ and $P$-value of 0.000, we conclude that the actual citations do not match the population distributions. It is clear from the above the the largest contributions come from the youngest and oldest age groups. The younger ones are cited much more than expected, the older ones much less.

**22.17** If births are equally spread throughout the year, each sign should have 1/12 of them. We have $H_0$: all signs have probability 1/12. $H_A$ is that $H_0$ is false. We will perform a $\chi^2$ goodness-of-fit test with the given data. (It is reasonable to assume the GSS is a random survey of all US adults.) The data given represent 4344 individuals. Under the null hypothesis, we expect 4344/12 = 362 individuals in each sign. We omit details (see Exercises 22.113 and 22.15 above), and find $\chi^2$ = 19.76 with $P$-value 0.049, barely significant at the 5% level. We reject $H_0$ and conclude births are not equally spread through the year. We can see that Aries and Virgo make the largest contributions to the statistic – Aries (a winter month) has a lower than expected count and Virgo (a fall month) has a higher than expected count.

**22.29** If we combine the races, we have 140 + 976 + 121 = 1237 individuals who would let the racist speak and 129 + 480 + 131 = 740 who would not, making a total sample of size $n$ = 1977. The observed proportion who would allow a racist to speak is $\hat{p} = 1237/1977 = 0.6257$.




| Low |
|---|
| 0.598 |



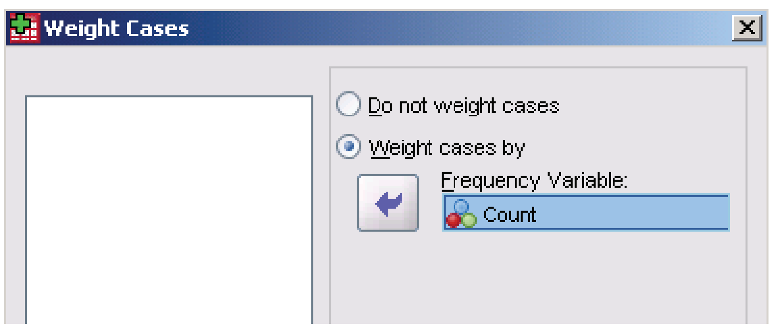

| High |
|---|
| 0.654 |

Based on this GSS survey, between 59.8% and 65.4% of U.S. adults think a racist should be allowed to speak, with 99% confidence.

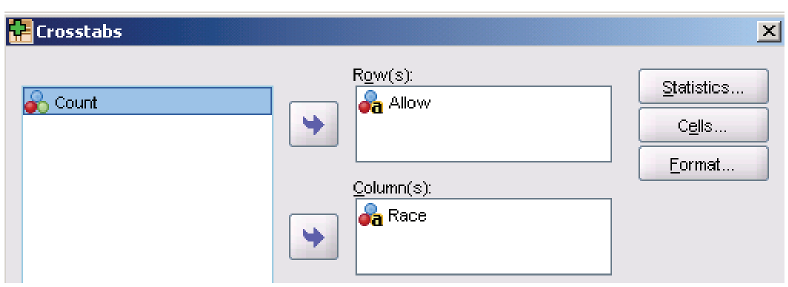
There were 269 Blacks, of whom 140/269 = 52.0% thought racists should be allowed to speak. For Whites, the percent is 976/(976+480) = 67.0%; for Others we have 121/252 = 48.0%. Both the Blacks and Others have percentages much less than Whites, but there were more Whites in the sample. To perform the chi-square test, enter the data as below.

| Race | Allow | Count |
|---|---|---|
| black | yes | 140 |
| black | no | 129 |
| white | yes | 976 |
| white | no | 480 |
| other | yes | 121 |
| other | no | 131 |

Click **<u>Data</u>**, **<u>Weight Cases</u>**. Click to weight cases by **Count**, then **<u>OK</u>**.




Now, click **<u>Analyze</u>**, **<u>Descriptive Statistics</u>**, **<u>Crosstabs</u>**. Click to enter **Allow** as the row and **Race** as the column. Now, click the **<u>Statistics</u>** button and check the box to ask for the Chi-square. **<u>Continue</u>** and click the **<u>Cells</u>** button. Click to ask for the observed and expected counts. **<u>Continue</u>** and **<u>OK</u>** computes the test.




We have the table below with both observed and expected counts.

**Allow * Race Crosstabulation**

| | | | Race | | | |
|---|---|---|---|---|---|---|
| | | | black | other | white | Total |
| Allow | no | Count | 129 | 131 | 480 | 740 |
| | | Expected Count | 100.7 | 94.3 | 545.0 | 740.0 |
| | yes | Count | 140 | 121 | 976 | 1237 |
| | | Expected Count | 168.3 | 157.7 | 911.0 | 1237.0 |
| | Total | Count | 269 | 252 | 1456 | 1977 |
| | | Expected Count | 269.0 | 252.0 | 1456.0 | 1977.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 47.899[a] | 2 | .000 |
| Likelihood Ratio | 46.952 | 2 | .000 |
| N of Valid Cases | 1977 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 94.32.

The *P*-value of the test is 0.000. We have overwhelming evidence that more whites would allow a racist to speak than Blacks or people of other ethnicities. Note that the largest contributions to the test statistic are from the Other column.

**22.43** We're using the data layout from file *ex22-43*. This file has race in a column, school opinion in one, and the counts in a third. We again use the variable Count to weight cases, then use **Analyze**, **Descriptive Statistics**, **Crosstabs** as described in Exercise 22.29.

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 22.426[a] | 8 | .004 |
| Likelihood Ratio | 22.897 | 8 | .003 |
| N of Valid Cases | 605 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 21.26.

**Schools * Race Crosstabulation**

| | | | Race | | | |
|---|---|---|---|---|---|---|
| | | | black | white | hispanic | Total |
| Schools | don't | Count | 22 | 14 | 28 | 64 |
| | | Expected Count | 21.4 | 21.3 | 21.4 | 64.0 |
| | excel | Count | 12 | 22 | 34 | 68 |
| | | Expected Count | 22.7 | 22.6 | 22.7 | 68.0 |
| | fair | Count | 75 | 60 | 61 | 196 |
| | | Expected Count | 65.4 | 65.1 | 65.4 | 196.0 |
| | good | Count | 69 | 81 | 55 | 205 |
| | | Expected Count | 68.4 | 68.1 | 68.4 | 205.0 |
| | poor | Count | 24 | 24 | 24 | 72 |
| | | Expected Count | 24.0 | 23.9 | 24.0 | 72.0 |
| | Total | Count | 202 | 201 | 202 | 605 |
| | | Expected Count | 202.0 | 201.0 | 202.0 | 605.0 |

The differences in the distributions are statistically significant ($P$ = 0.004). To see the departures from the null hypothesis, examine the expected counts. Blacks are less likely to call schools Excellent than expected (12 observed versus 22.7 expected) while Hispanics are more likely to call them Excellent (34 observed and 22.7 expected) and less likely to call them Good (55 versus 68). Blacks are more likely to call them Good (75 versus 65.4). There seems to be no real differences among the ethnicities on calling the schools Poor.

**22.45** We've used the data in *ex22-45*. As in the last two exercises, we use Data, Weight Cases to weight the results by Count. We then use **Analyze**, **Descriptive Statistics**, **Crosstabs** to recreate the table and add the expected counts (click **Cells**, **Expected**).

**Newpref * Group Crosstabulation**

| | | | Group | | | | |
|---|---|---|---|---|---|---|---|
| | | | hardhot | hardwarm | softhot | softwarm | Total |
| Newpref | no | Count | 30 | 42 | 27 | 53 | 152 |
| | | Expected Count | 30.9 | 47.2 | 24.0 | 49.8 | 152.0 |
| | yes | Count | 42 | 68 | 29 | 63 | 202 |
| | | Expected Count | 41.1 | 62.8 | 32.0 | 66.2 | 202.0 |
| | Total | Count | 72 | 110 | 56 | 116 | 354 |
| | | Expected Count | 72.0 | 110.0 | 56.0 | 116.0 | 354.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 2.058[a] | 3 | .560 |
| Likelihood Ratio | 2.062 | 3 | .560 |
| N of Valid Cases | 354 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 24.05.

There is no significant difference between the person's laundry practice and their preference for the new product ($P = 0.560$), although it appears that the people with soft water seem to prefer the standard product (their expected counts are somewhat smaller than the observed) and the people with hard water seem to prefer the new product (their expected counts are also a bit smaller than observed).

**22.47** The new table will be as shown below.

| | None | High School | Jr. college | Bachelor | Graduate |
|---|---|---|---|---|---|
| Democrat leaning | 279 | 996 | 156 | 313 | 218 |
| Republican leaning | 135 | 731 | 129 | 336 | 128 |

To see if support differs by level of education, we enter the data as shown below. As in the last exercises, we weight cases by Count and use Analyze, Descriptive Statistics, Crosstabs to compute the test. Do not forget to ask for the Chi-squared **Statistic** and the **Cell** **Expected** values.

| Leaning | Education | Count |
|---|---|---|
| Democrat | None | 279 |
| Democrat | HS | 996 |
| Democrat | JC | 156 |
| Democrat | Bachelor | 313 |
| Democrat | Graduate | 218 |
| Republican | None | 135 |
| Republican | HS | 731 |
| Republican | JC | 129 |
| Republican | Bachelor | 336 |
| Republican | Graduate | 128 |

**Leaning * Education Crosstabulation**

| | | | Education | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Bachelor | Graduate | HS | JC | None | Total |
| Leaning | Democrat | Count | 313 | 218 | 996 | 156 | 279 | 1962 |
| | | Expected Count | 372.2 | 198.4 | 990.5 | 163.5 | 237.4 | 1962.0 |
| | Republican | Count | 336 | 128 | 731 | 129 | 135 | 1459 |
| | | Expected Count | 276.8 | 147.6 | 736.5 | 121.5 | 176.6 | 1459.0 |
| | Total | Count | 649 | 346 | 1727 | 285 | 414 | 3421 |
| | | Expected Count | 649.0 | 346.0 | 1727.0 | 285.0 | 414.0 | 3421.0 |

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 44.539[a] | 4 | .000 |
| Likelihood Ratio | 44.806 | 4 | .000 |
| N of Valid Cases | 3421 | | |

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 121.55.

With a 0.000 *P*-value, we conclude there is a difference in political leaning with education level. People with no high school education are more likely to lean Democrat as are people with either a Bachelor's or graduate degree; in other words, the Democrats seem to draw support from either people with little or a lot of education.