

SPSS Lesson 3

A Few Basic Statistics

Your statistical analyses are not the most important part of the research report you will be writing, but you will be expected to be able to (1) generate frequency distributions and basic descriptive statistics for your key variables; (2) examine relationships among variables by (a) looking at associations between two variables by way of cross-tabulations and the chi-square statistic; and/or (b) looking at differences between two groups via the t-test.

FREQUENCIES

The FREQUENCIES procedure in SPSS offers the simplest way to get both frequency distributions and summary statistics for each of your variables. Let's walk through an example from the QuickSurvey database.

- Start SPSS and open up the QuickSurvey database.
- Click ANALYZE on the top menu, then DESCRIPTIVE STATISTICS on the drop down menu, and FREQUENCIES on the menu after that.
- You see a FREQUENCIES dialogue box before you. On the left hand side is a list of all the variables. Click "Where were you born?" [WHERBORN], and then click the arrow, which will bring that variable in the box to the right entitled "VARIABLE(S):"
- "DISPLAY FREQUENCY TABLES" should be checked.
- Click on "STATISTICS" at the bottom of the dialogue box, and another box will open in which you place your order with SPSS from the statistics menu. It is up to you to make some wise choices here about what sorts of statistics are meaningful to request. A "typical" assortment would include MEAN, MEDIAN, MODE and STANDARD DEVIATION. Let's click those and then click CONTINUE.
- We are back at the FREQUENCIES dialogue box. If you wanted to, you could now click CHARTS and go to that dialogue box to order graphs of various sorts, but we'll leave that discussion for next week. The FORMAT dialogue box allows you to specify what order you would like the frequencies specified in (e.g., by ascending or descending codes; by ascending or descending frequencies).
- Press OK and a new window opens that is called OUTPUT. You can save the products that emerge in this OUTPUT window as a separate file.
- When I pressed OK after setting up a FREQUENCIES analysis of WHERBORN, I got the following:

Statistics

Where were you born?

N	Valid	92
	Missing	1
Mean		2.20
Median		2.00
Mode		1
Std. Deviation		1.42

- When I said that it was up to you to make wise choices about what statistics to generate, I meant it, and the statistics I've produced here are a case in point. Note that WHERBORN is a nominal/categorical variable, with the different levels in the coding reflecting differing degrees of distance from the Lower Mainland area of British Columbia. We might go so far as to call it an Ordinal level variable because of the order inherent to the categories. But does it make any sense to ask for the mean (average) category? No. But did SPSS produce this statistic when I asked for it? Yes – you can see the mean is 2.20. SPSS is not smart enough to recognize when we are asking for stupid things – it will conjure up statistics no matter what we ask for -- so it is up to us to determine what is stupid and what is meaningful.
- The only meaningful statistic here would be the mode, which shows that code 1 – persons who said they were born in the Lower Mainland area of BC – were the most frequently occurring category.

Where were you born?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Lower Mainland	43	46.2	46.7	46.7
	Elsewhere in BC	16	17.2	17.4	64.1
	Elsewhere in Canada	19	20.4	20.7	84.8
	Elsewhere in the World	14	15.1	15.2	100.0
	Total	92	98.9	100.0	
Missing	9	1	1.1		
Total		93	100.0		

- As for the frequency distribution portion of the output, note that because we went through the variable definition process outlined in SPSS Lesson 1, the code-by-code value labels are already listed.
- The “FREQUENCY” column shows you the number of people who checked each alternative. Note that 92 people answered the question; 1 did not.
- There are three “percent” columns – the first simply says “percent,” while the second says “valid percent.” The difference between these two is that “percent” takes the number of respondents as a percentage of all participants – including the people who did *not* answer – while “valid percent” expresses the number responding as a percent

of those who responded. Normally, the “valid percent” is the one of these two that you want, but there are some occasions where you might make the opposite choice. [As an aside, I’ll note that using the word “valid” here in SPSS is a poor choice on their part; they mean “valid” simply in the sense that the person answered the question; it is not an endorsement of its psychometric validity.] “Cumulative percent” is rarely utilized.

CROSS-TABULATIONS AND CHI-SQUARE

Cross-tabulations allow you to look at how changes in the frequency of occurrence of one variable is associated with changes in the frequency of occurrence of another variable. It is ideally suited when the two variables are both nominal (i.e., categorical), but can be used to look at the association of any two variables, as long as the number of categories is not terribly large, and there are not too many empty cells.

Chi-square is a statistic that is often used to arbitrate whether the degree of association we observe between the two variables is greater than what we’d expect in the basis of chance alone, i.e., whether the association is statistically significant or not.

Let’s do one using the QuickSurvey database:

- Start SPSS and open up the QuickSurvey database.
- The first thing you have to decide is what two variables you want to cross-tabulate, and then how you would like these two to appear in the cross-tabulation (i.e., which will be the row variable and which will be the column variable). I would like to cross-tabulate “Gender” with having participated in a “Team-Based sport.” I will make Gender the row variable and Participation in a team-based sport the column variable.
- Click ANALYZE on the top menu, then DESCRIPTIVE STATISTICS on the drop down menu, and CROSSTABS on the menu after that.
- You see a CROSSTABS dialogue box before you. On the left hand side is a list of all the variables. Click “Gender of Respondent” [GENDER] and then click the arrow to bring that variable into the “ROW(S)” box; then click on “Played team-based sport” and then click on the arrow to take that variable into the box marked “COLUMN(S).”
- At the bottom of the dialogue box is a button that says “STATISTICS.” Click that.
- A “STATISTICS” dialogue box opens. You have many options, but the basic for me is that you choose “CHI-SQUARE.” Then click CONTINUE.
- Back at the CROSSTABS dialogue box, you also see a button at the bottom labelled CELLS. Click that, and you now see a CROSSTABS: CELL DISPLAY dialogue box. The minimum here to check is “OBSERVED” (under “Counts”), and then whichever (row or column) percentage breakdown you are interested in. Because GENDER is my row variable, and because I want to know what percentage of men and women engage in team-based sporting activities, I check “ROW” here. To close the dialogue box click CONTINUE.
- Back at the CROSSTABS dialogue box, if everything looks OK to you, press OK.

- When I did, I got the following output:
- The first little table, which you see below, is nothing more than a statement that there were 93 cases on which the analysis is based, i.e., 93 people for whom information on both variables was available. No cases were missing.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender of Respondent * Played team-based sport	93	100.0%	0	.0%	93	100.0%

- The next table gives the cross tabulation we asked for, in the form we asked it to be presented in.

Gender of Respondent * Played team-based sport Crosstabulation

			Played team-based sport		Total
			No	Yes	
Gender of Respondent	Men	Count % within Gender of Respondent	19 55.9%	15 44.1%	34 100.0%
	Women	Count % within Gender of Respondent	49 83.1%	10 16.9%	59 100.0%
Total		Count % within Gender of Respondent	68 73.1%	25 26.9%	93 100.0%

- Next (see below) comes a list of assorted chi-square tests, most of which are of no interest to us. The one you should pay attention to is the first one, i.e., the Pearson Chi-Square statistic.
- Reading across that row you see the chi-square value is 8.100 with 1 degree of freedom. The next column notes that the probability of getting this sort of result on the basis of chance variation alone is .004 (i.e., 4 out of 1000). Because this is less than .05 (5 out of 100), we say that the result is statistically significant, i.e., we conclude that gender and involvement in team-based sports are related.
- Inspection of the table above reveals that this is because the men are about three times more likely than the women to say they participated in team-based sports (44.1% vs 16.9%).

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	8.100 ^b	1	.004		
Continuity Correction ^a	6.777	1	.009		
Likelihood Ratio	7.905	1	.005		
Fisher's Exact Test				.007	.005
Linear-by-Linear Association	8.013	1	.005		
N of Valid Cases	93				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.14.

LOOKING AT DIFFERENCE VIA THE T-TEST

T-tests allow you to look at differences in the means of two groups. For example, in the QuickSurvey database, there were two versions of the survey circulated. The difference was with respect to the four Likert-type items that appeared at the bottom of the first page. For version “A” of the survey, the first three items alluded to problems within Canada that could be used to promote a more protectionist stance; for version “B,” the first three items alluded to Canada’s positive international image and its willingness to do its part in the world community of nations. The big question was whether these two set-ups would result in different responses to the immigration items that appeared as the fourth item in both versions. That is a job for – the T-TEST!

- Start SPSS and open up the QuickSurvey database.
- Click ANALYZE on the top menu, then COMPARE MEANS on the drop down menu, and INDEPENDENT SAMPLES T-TEST on the menu after that.
- A dialogue box opens. As usual, there is a list of variables on the left. On the right are two boxes, entitled “Test Variable(s)” and “Grouping Variable.” The “Grouping Variable” refers to the variable that contains the two groups you want to compare; in this case the grouping variable is VERSION, i.e., the two different versions of the survey. The “Test Variable(s)” refers to the dependent variable, which, in this case is responses on the “immigration” item. Click on the appropriate variable name and then the appropriate arrow to get the two variables in their respective boxes.
- Note that the grouping variable has two question marks in parentheses, and a button underneath entitled DEFINE GROUPS. That’s where we go next, because SPSS wants us to be clear on the identity of the two groups we want to compare. So ... click DEFINE GROUPS. A dialogue box opens in which you put the codes for the two groups. In our case this is simply A and B, which is how that variable was coded. When you’re finished, click CONTINUE.
- You’re back an the T-Test dialogue box; press OK. A new OUTPUT page appears. Mine contains the following information:

Group Statistics

Version of Questionnaire Completed		N	Mean	Std. Deviation	Std. Error Mean
Immigration levels too high	Protectionist	42	2.83	1.53	.24
	Internationalist	50	2.74	1.08	.15

- You can see that 42 people received the “protectionist” version of the survey. And 50 the “internationalist” one.
- You also can see that the means for the two groups were 2.83 and 2.74 on the 5-point scale where 1 = disagree strongly and 5 = agree strongly.
- This means that, as hypothesized, attitudes about immigration were more positive in the group who received the internationalist version of the survey.
- But was the difference significant? We have to check the next table that’s entitled “Independent Samples Test.” In most cases, the line you will want to look at is the top line where it says “test for equality of means.” This reveals that the t-statistic is less than 1 (.341 to be exact) which, for 90 degrees of freedom, is clearly non-significant (the exact probability given is .734).

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference
Immigration levels too high	Equal variances assumed	9.692	.002	.341	90	.734	9.33E-02
	Equal variances not assumed			.332	72.188	.741	9.33E-02