Supplement to Chapter 17

SPSS Analysis of Descriptive Statistics

Computer software is almost invariably used to compute descriptive statistics. There are several widely used software packages for statistical analysis. The one we used to generate some examples of statistical output is called Statistical Package for the Social Sciences (SPSS).[1] (Because IBM acquired SPSS several years ago, the official name for the software is IBM SPSS Statistics.)

DATA FOR THE SPSS EXAMPLE

To illustrate descriptive statistics from SPSS, we will use a very small, fictitious dataset. Let us suppose that we were evaluating the effectiveness of a complex intervention for low-income pregnant adolescents. The intervention is a program of health care, nutrition education, contraceptive counseling, and parenting education. Thirty young women are randomly assigned to either the special program (the experimental group) or usual care (the control group). Two key outcomes are infant birthweight and a repeat pregnancy within 18 months of delivery.

For an analysis in SPSS, data are entered into a data file that is arrayed like a spreadsheet. Each row represents a study participant, and every column represents a study variable. Figure 1 shows a screenshot of an SPSS data file, with fictitious data for our example. The first column is simply an identification number (ID). The next column (GROUP) indicates whether a participant is in the intervention group (coded 1) or in the "usual care" group (coded 2). The next three columns contain data about three maternal characteristics: the mothers' age (AGE), number of prior pregnancies (PRIORS), and whether or not they smoke (SMOKE: 1 = smokes, 0 = does not smoke). The next column shows the infant's birthweight, in ounces (BWEIGHT). Finally, the last column shows the data for whether or not the mother had a repeat pregnancy within 18 months after giving birth (REPEAT: 1 = yes, 0 = no).

This dataset includes three variables measured on a nominal scale: GROUP, SMOKE, and REPEAT. Both AGE and BWEIGHT are ratio-level variables. For the *raw data* in the dataset (data before any manipulations), there are no ordinal- or interval-level variables. Now we can illustrate how some descriptive statistics for this dataset can be produced in SPSS.

FREQUENCY DISTRIBUTION

A good place to begin any analysis is to look at the frequency distributions for all variables. The SPSS commands for this, which can be selected from menus, are:

ANALYZE → DESCRIPTIVE STATISTICS → FREQUENCIES

[1]SPSS Version 21 was used to create all output.

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons

	ID	GROUP	AGE	PRIORS	SMOKE	BWEIGHT	REPEAT
1	1	1	17	1	1	107	1
2	2	1	14	0	0	101	0
3	3	1	21	3	0	119	0
4	4	1	20	2	0	128	1
	5	1	15	1	1	89	0
6	6	1	19	0	1	99	0
7	7	1	19	1	0	111	0
8	8	1	18	1	1	117	1
9	9	1	17	0	0	102	1
	10	1	20	0	0	120	0
11	11	1	13	0	1	76	0
12	12	1	18	0	1	116	0
13	13	1	16	0	0	100	1
14	14	1	18	0	0	115	0
	15	1	21	2	1	113	0
16	16	2	19	0	0	111	1
17	17	2	21	1	0	108	0
18	18	2	19	2	1	95	0
19	19	2	17	0	1	99	0
	20	2	19	0	0	103	1
21	21	2	15	0	1	94	0
22	22	2	17	1	0	101	1
23	23	2	21	2	0	114	0
24	24	2	20	1	0	97	0
	25	2	18	0	1	99	1
26	26	2	18	0	1	113	0
27	27	2	19	1	0	89	0
28	28	2	20	0	0	98	0
29	29	2	17	0	0	102	0
	30	2	19	1	1	105	0

NOTES:

GROUP:	Group status,	1 = Experimental group	2 = Control group
AGE:	Mother's age in years		
PRIORS:	Number of prior pregnancies		
SMOKE:	Mother's smoking status,	1 = Smokes	0 = Does not smoke
BWEIGHT:	Infant's birthweight, in ounces		
REPEAT:	Had repeat pregnancy within 18 months,	1 = Yes	0 = No

FIGURE 1 Fictitious dataset for intervention study with low-income pregnant adolescents (screenshot of an SPSS Data File).

Figure 2 presents the SPSS FREQUENCY output for the variable BWEIGHT (infant birthweight). Panel A shows several descriptive statistics. The *Mean* is 104.70, the *Median* is 102.50, and the *Mode* is 99.00, suggesting a modestly skewed distribution. (SPSS can also compute a skewness index, an index that we did not explain in the textbook. The skewness statistic for this variable was −.254, a mild negative skew.) The standard deviation for BWEIGHT (*Std Deviation*) is 10.95, and the *Variance* is 120.01 (10.95^2). The *Range* is 52.00, which is equal to the *Maximum* of 128.00 minus the *Minimum* of 76.00.

The frequency distribution for BWEIGHT is shown in panel B of Figure 2. Each birthweight is listed in the first column, from the low value of 76 to the high value of 128. The next column, *Frequency*, shows the number of occurrences of each birthweight. There was one 76-ounce baby, two 89-ounce babies, and so on. The next column, *Percent*, indicates the percentage of infants in each birthweight category: 3.3% weighed 76 ounces, 6.7% weighed 89 ounces, and so on. The next column, *Valid Percent*, indicates the percentage in each category after removing any missing values from the denominator. In this example, birthweights were obtained for all 30 infants, but if one birthweight had been missing, the valid percent for the 76-ounce baby would have been 3.4% (1 ÷ 29 rather than 1 ÷ 30). The last column, *Cumulative Percent*, adds the percentage for a given birthweight to the percentage for all preceding values. Thus, we can tell by looking at the shaded row for 99 ounces that, cumulatively, 33.3% of the babies weighed less than 100 ounces.

When variables are continuous with many different values, it may be cumbersome to examine a complete frequency distribution showing counts and percentages. For example, if there were 500 sample members, there might be a birthweight for every value between 76 and 128 (or more). In such cases, we can instruct SPSS to print only panel A, the central tendency and variability information. Alternatively, we can use the following command:

ANALYZE → DESCRIPTIVE STATISTICS → DESCRIPTIVES

When we used this command to calculate descriptive statistics for both BWEIGHT and AGE, we get the results shown in Figure 3. This is an efficient way of getting descriptive statistics for many continuous variables all at once. Note that another SPSS program that provides descriptive information for individual variables is EXPLORE, which we discuss in the Supplement to Chapter 20.

HISTOGRAM

SPSS also can create graphic displays of the data for individual variables. For example, histograms can be generated within the FREQUENCIES routine, the EXPLORE routine, or with a separate GRAPHS command.

Figure 4 shows a histogram for maternal age, which we created in the FREQUENCIES program. The age values (ranging from 13 to 21 years) are on the horizontal axis, and frequencies are on the vertical axis. The histogram shows at a glance that the modal age is 19 years ($f = 7$) and that age is negatively skewed (i.e., there are few very young mothers). Descriptive statistics to the left of the histogram indicate that the mean age for this group is 18.17, with an SD of 2.09—the same values that are in Figure 3, but rounded to two decimal places.

CROSSTABULATIONS

SPSS can also be used to compute bivariate descriptive statistics, such as crosstabs and correlations. In the current example, we will compare the repeat pregnancy rate of experimental versus control group mothers. Both of these variables (REPEAT and GROUP) are nominal-level measures, and so we used the following SPSS commands:

ANALYZE → DESCRIPTIVE STATISTICS → CROSSTABS

The results are presented in the crosstabs table in Figure 5. The crosstab analysis resulted in four main cells: (1) intervention (experimental) group mothers with no repeat pregnancy (upper left cell), (2) control group mothers with no repeat pregnancy, (3) intervention group mothers with a repeat pregnancy, and (4) control group mothers with a repeat pregnancy.

Each cell in the crosstabs table contains four pieces of information. The first is number of people in the cell (*Count*). In the first cell (shaded in blue), 10 experimental group participants did not

A **Statistics**

Infant birth weight in ounces

N	Valid	30.00
	Missing	.00
	Mean	104.70
	Median	102.50
	Mode	99.00
	Std. deviation	10.95
	Variance	120.01
	Range	52.00
	Minimum	76.00
	Maximum	128.00

B **Infant birth weight in ounces**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	76	1	3.3	3.4	3.3
	89	2	6.7	6.7	10.0
	94	1	3.3	3.3	13.3
	95	1	3.3	3.3	16.7
	97	1	3.3	3.3	20.0
	98	1	3.3	3.3	23.3
	99	3	10.0	10.0	33.3
	100	1	3.3	3.3	36.7
	101	2	6.7	6.7	43.3
	102	2	6.7	6.7	50.0
	103	1	3.3	3.3	53.3
	105	1	3.3	3.3	56.7
	107	1	3.3	3.3	60.0
	108	1	3.3	3.3	63.3
	111	2	6.7	6.7	70.0
	113	2	6.7	6.7	76.7
	114	1	3.3	3.3	80.0
	115	1	3.3	3.3	83.3
	116	1	3.3	3.3	86.7
	117	1	3.3	3.3	90.0
	119	1	3.3	3.3	93.3
	120	1	3.3	3.3	96.7
	128	1	3.3	3.3	100.0
	Total	30	100.0	100.0	

FIGURE 2 SPSS printout of a frequency distribution for infant birthweight (BWEIGHT).

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Infant birth weight in ounces	30	52.00	76.00	128.00	104.7000	10.95492	120.010
Mother's age	30	8.00	13.00	21.00	18.1667	2.08580	4.351
Valid *N* (listwise)	30						

FIGURE 3 SPSS printout of descriptives for BWEIGHT and AGE.

Frequencies

Statistics

Mother's age

N	Valid	30.00
	Missing	.00
	Mean	18.17
	Median	18.50
	Mode	19.00
	Std. deviation	2.09
	Variance	4.35
	Range	8.00
	Minimum	13.00
	Maximum	21.00

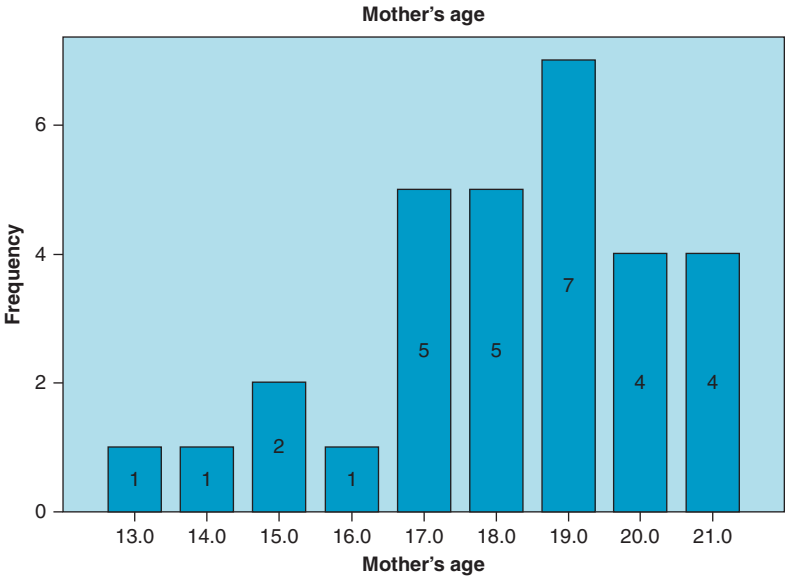

FIGURE 4 SPSS printout: frequencies and histogram for maternal age (AGE).

have a repeat pregnancy within 18 months of delivery. Below the 10 is the *row* percentage or *% within Repeat pregnancy*: 47.6% of the women who did not become pregnant again were in the intervention group (10 ÷ 21). The next entry in the cell is the *column* percentage or *% within Treatment group*: 66.7% of the experimental group mothers did not have a repeat pregnancy (10 ÷ 15). Last is the *overall* percentage of participants who were in that cell or *% of Total* (10 ÷ 30 = 33.3%). Figure 5 indicates that a somewhat higher percentage of experimental group (33.3%) than control group (26.7%) participants had an early repeat pregnancy, as shown in the shaded row. The row totals on the far right indicate that, overall, 30.0% of the sample ($N = 9$) had a subsequent pregnancy. The column totals at the bottom indicate that 50.0% of all participants were in the control group and 50.0% were in the experimental group.

If the results from this table were presented in a report, only the column percentages likely would be shown because the column percentages add to 100%. For example, for the control group, 73.3% (no repeat pregnancy) and 26.7% (a repeat pregnancy) add to 100.0% (as shown in the shaded values in the control group column of Figure 5).

RISK INDEXES

SPSS can calculate the odds ratio and relative risk index within the CROSSTABS program. (In the menu for statistics, click on "Risk.") For the

Repeat pregnancy * Treatment group Crosstabulation

			Treatment group		Total
			Experimental	Control	
Repeat pregnancy	No	Count	10	11	21
		% within repeat pregnancy	47.6%	52.4%	100.0%
		% within treatment group	66.7%	73.3%	70.0%
		% of total	33.3%	36.7%	70.0%
	Yes	Count	5	4	9
		% within repeat pregnancy	55.6%	44.4%	100.0%
		% within treatment group	33.3%	26.7%	30.0%
		% of total	16.7%	13.3%	30.0%
Total		Count	15	15	30
		% within repeat pregnancy	50.0%	50.0%	100.0%
		% within treatment group	100.0%	100.0%	100.0%
		% of total	50.0%	50.0%	100.0%

FIGURE 5 SPSS printout: crosstabulation of repeat pregnancy and treatment group status.

analysis of risk for the experimental versus control group having a repeat pregnancy, we reversed the arrangement of the crosstabs table. That is, we put GROUP in the rows and REPEAT in the columns, so that the data would be arrayed as in Table 16.6 of the textbook. We do not show the actual reconfigured crosstabs table, but Figure 6 presents the output for the risk information. (Ignore the confidence interval information in the two right-most columns—confidence intervals are explained in Chapter 17.)

Figure 6 shows that the odds ratio (OR) for group status is 1.375 (shaded). The estimated odds of having a repeat pregnancy are about 38% higher for those in the intervention group. The next row in Figure 5 shows the relative risk (RR) for having a repeat pregnancy—although the output does not make this clear. Given the status of being in the experimental group, the RR is 1.250 (shaded). The next row can be ignored—it represents the relative risk of *not* having a repeat pregnancy, given the status of being in the intervention group.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds ratio for treatment group (experimental/control)	1.375	.286	6.603
For cohort repeat pregnancy = Yes	1.250	.415	3.766
For cohort repeat pregnancy = No	.909	.568	1.455
N of valid cases	30		

FIGURE 6 SPSS printout: odds ratio and relative risk for repeat pregnancy and treatment group status.