

**Abstract**

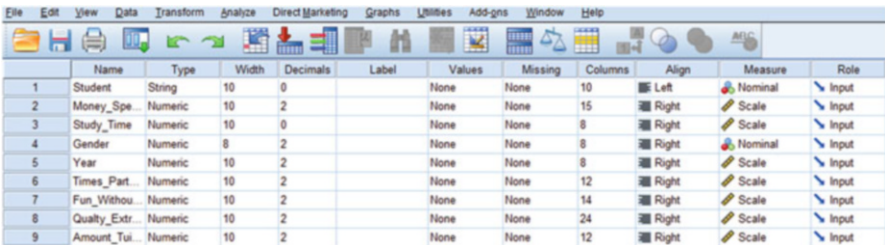
This first practical chapter is split into two parts. In the first part, I succinctly present the two statistical software packages, SPSS and Stata, which are probably the most used statistical programs in the social sciences. I also explain how to input data into SPSS and Stata datasets. Second, I cover the two univariate statistical categories, frequency tables and descriptive statistics, as well as some graphical representations of a variable (i.e., boxplot, pie chart, and histogram). For each test, the mathematical foundations as well as the practical implementation in SPSS and Stata will be introduced based on the sample survey presented at the end of Chap. 4.

**6.1 SPSS and Stata**

SPSS and Stata are probably the most frequently used software packages in introductory statistics' classes. Both programs are complete, integrated statistics packages that allow for data analysis, data management, and graphics. They are available in most university libraries and are rather simple to navigate. For each test or graphic, the book will illustrate the logic behind the test, its mathematical foundations, as well as illustrate how to conduct the respective analysis in both SPSS and Stata. Depending on which software package you use, you only need to read either the SPSS or the Stata example.

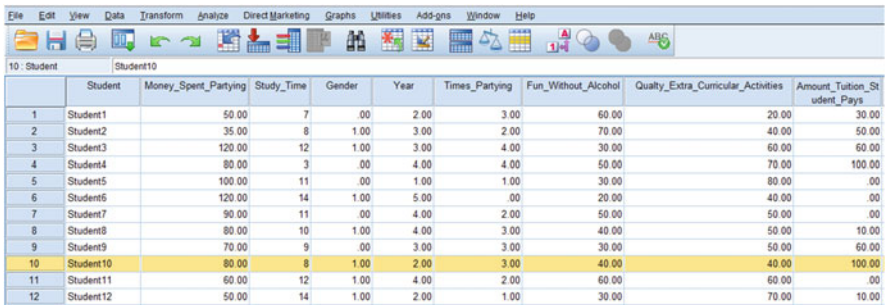
**6.2 Putting Data into an SPSS Spreadsheet**

To create a data file in SPSS, open SPSS then click on *new dataset*. This opens up a spreadsheet similar to an Excel document. There are two parts to any SPSS dataset, one part labeled Variable View and one part labeled Data View. Data View is used to enter data. Variable View is used to set up the data—names, variable labels, value



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Student	String	10	0		None	None	10	Left	Nominal	Input
2	Money_Spe...	Numeric	10	2		None	None	15	Right	Scale	Input
3	Study_Time	Numeric	10	0		None	None	8	Right	Scale	Input
4	Gender	Numeric	8	2		None	None	8	Right	Nominal	Input
5	Year	Numeric	10	2		None	None	8	Right	Scale	Input
6	Times_Part...	Numeric	10	2		None	None	12	Right	Scale	Input
7	Fun_Withou...	Numeric	10	2		None	None	14	Right	Scale	Input
8	Quality_Ext...	Numeric	10	2		None	None	24	Right	Scale	Input
9	Amount_Tui...	Numeric	10	2		None	None	12	Right	Scale	Input

Fig. 6.1 Display of the sample data in Variable View



	Student	Money_Spent_Partying	Study_Time	Gender	Year	Times_Partying	Fun_Without_Alcohol	Quality_Extra_Curricular_Activities	Amount_Tuition_Student_Pays
1	Student1	50.00	7	.00	2.00	3.00	60.00	20.00	30.00
2	Student2	35.00	8	1.00	3.00	2.00	70.00	40.00	50.00
3	Student3	120.00	12	1.00	3.00	4.00	30.00	60.00	60.00
4	Student4	80.00	3	.00	4.00	4.00	50.00	70.00	100.00
5	Student5	100.00	11	.00	1.00	1.00	30.00	80.00	.00
6	Student6	120.00	14	1.00	5.00	.00	20.00	40.00	.00
7	Student7	90.00	11	.00	4.00	2.00	50.00	50.00	.00
8	Student8	80.00	10	1.00	4.00	3.00	40.00	50.00	10.00
9	Student9	70.00	9	.00	3.00	3.00	30.00	50.00	60.00
10	Student10	80.00	8	1.00	2.00	3.00	40.00	40.00	100.00
11	Student11	60.00	12	1.00	4.00	2.00	60.00	60.00	.00
12	Student12	50.00	14	1.00	2.00	1.00	30.00	70.00	10.00

Fig. 6.2 Display of the sample data in Data View

labels, etc. To change from *Data View* to *Variable View*, click on the icons in the lower left corner of the dataset.


The first step of creating a dataset consists normally of labeling the data. We generally do this labeling in the *Variable View*. In the *Variable View*, each row represents one variable. Each column represents a case such as a person or a country (see Fig. 6.1). The first variable we normally enter into a dataset is a string variable, an identifier of the cases for which we have collected data for. To enter your survey data, go to the *Variable View* and write in “respondent” in the first row of the first column labeled *Name*. If you enter your own data, label these data appropriately. Because this first variable is normally a string variable (i.e., a non-numeric variable), choose the option *String* in the second column. The other variables are normally numeric variables—the actual data. Include their variable names in the subsequent rows, and make sure that these variables are labeled as *Numeric* (i.e., check the second column). When you label the data, make sure that you do not leave any space between letters, as SPSS will not allow you to leave any space.

The copied graph above shows the *Variable View* of the data from our sample questionnaire from Sect. 4.11. The first variable, labeled student, is the identifier variable. The second variable is the dependent variable measuring the amount of money students spend per week while going out. The remaining variables are the independent variables.

Figure 6.2 shows the *Data View* of our sample questionnaire data. The variable in the first column is the identifier. The second column is the dependent variable, the

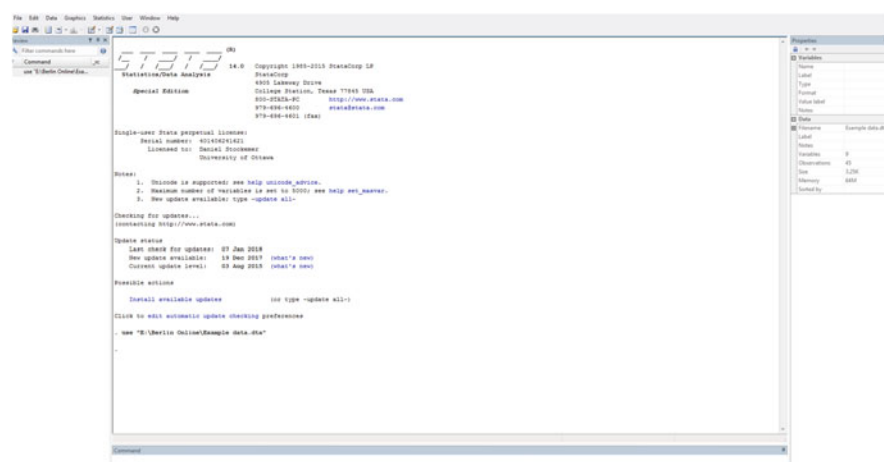
amount of money students spend going out. Variables 3–8 are the independent variables.

## 6.3 Putting Data into a Stata Spreadsheet

To create a data file, you have to click on the icon displaying a spreadsheet and a pencil  or type “edit” into the command line. This opens up a spreadsheet similar to an Excel document. You can then input the data by hand into this data editor. To do so, you have to first label the data. The first variable we normally enter into a dataset is a string variable, an identifier of the cases for which we have collected data. In our case, this is a string variable, which we label student. Add the name of the other variables in the subsequent fields of the first row. When you label the data, make sure that you do not leave any space between letters, as Stata does not allow you to leave any space. After labeling the variables, you can then input the data. Alternatively, you can enter the data in an Excel spreadsheet and copy it to Stata. When you paste the data into the Stata data editor, Stata asks you whether you want to treat the first row as variable names or data. You click variable names; otherwise, the data will not be labeled.

In Stata we have two different screens, the main screen we use to do our data analysis (see Fig. 6.3) and the data editor or the screen in which we can see the data (see Fig. 6.4). On the main screen, there is some information about Stata and an indication of the dataset you use. On the right side of the screen (i.e., the right upper corner), there is a listing of all the variables. On the bottom of the screen is the command editor, which you will use to do your analyses.

Figure 6.4 shows the data editor featuring our sample questionnaire data. The variable in the first column is the identifier. The second column is the dependent



**Fig. 6.3** The main Stata screen

	Student	Money_Spen-g	Study_Time	Gender	Year	Times_Part-g	Fun_Withou-1	Quality_Ex-v	Amount_Tui-s
1	Student1	50	7	0	2	2	60	90	30
2	Student2	35	8	1	3	1	70	40	50
3	Student3	120	12	1	3	4	30	20	60
4	Student4	80	3	0	4	4	50	50	100
5	Student5	100	11	0	1	1	30	10	0
6	Student6	120	14	1	5	4	20	20	0
7	Student7	90	11	0	4	2	50	50	0
8	Student8	80	10	1	4	3	40	50	10
9	Student9	70	9	0	3	3	30	50	60
10	Student10	80	8	1	2	3	40	40	100
11	Student11	60	12	1	4	2	60	40	0
12	Student12	50	14	1	2	1	30	70	10
13	Student13	100	13	0	3	4	0	30	0

**Fig. 6.4** Display of the sample data in the Stata data editor

variable, the amount of time students spend per week when partying. Variables 3–8 are the independent variables.

(Please note that when you create your Stata file, some of the original variable names might be too long, and Stata might not accept them. If this is the case, you have to shorten the original name.)

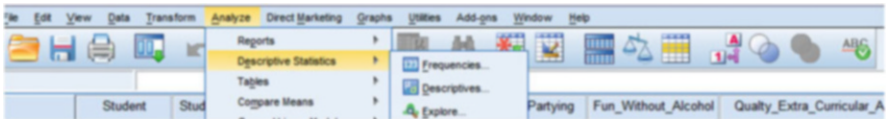
### 6.4 Frequency Tables

A frequency table is a rather simple univariate statistic that is particularly useful for categorical variables such as ordinal variables that do not have too many categories. For each category or value, such a table indicates the number of times or percentage of times each value occurs. A frequency table normally has four columns. The first column lists the available categories. The second column displays the raw frequency or the number of times a single value occurs. The third column shows the percentage of observations that fall into each category—the basic formula to calculate this percentage is the number of observations in the category divided by the total number of observations. The final column, labeled cumulative percentage, displays the percentage of individuals up to a certain category or point.

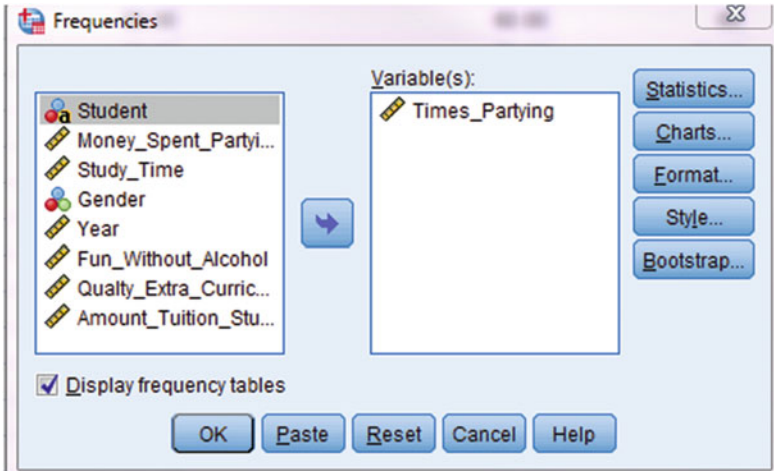
Table 6.1 displays a frequency table, of the variable *times partying*. The first column displays the available options (i.e., the six categories ranging from zero to five times and more). The second column shows the raw frequencies (i.e., how many individuals normally party, zero times, once, twice, three times, four times or five times, and more per week). The third column displays the raw frequency in percentages. The final column displays the cumulative percentage. For example, Table 6.1 highlights that 60% of the polled, on average, party two times or fewer per week.

**Table 6.1** Frequency table of the variable times partying

On average, how many times per week do you party	Frequency	Percentage	Cumulative percentage
Zero times	3	7.5	7.5
Once	8	20	27.5
Twice	13	32.5	60
Three times	10	25	85
Four times	5	12.5	97.5
Five or more times	1	2.5	100
Total	40	100	



**Fig. 6.5** Doing a frequency table in SPSS (first step)



**Fig. 6.6** Doing a frequency table in SPSS (second step)


**6.4.1 Constructing a Frequency Table in SPSS**

- Step 1: Go to Analyze—Descriptive Statistics—Frequencies (see Fig. 6.5).
- Step 2: Highlight the variable—times\_partying—and then click on the arrow. The variable appears in the right rectangle. Then click okay (see Fig. 6.6).

The SPSS output has five columns (see Table 6.2). Columns one, two, three, and five mimic Table 4.6. The first variable is the identifier and displays the existing

**Table 6.2** SPSS Frequency table output

Times_Partying		Frequency	Percent	Valid percent	Cumulative percent
Valid	0.00	3	7.5	7.5	7.5
	1.00	8	20.0	20.0	27.5
	2.00	13	32.5	32.5	60.0
	3.00	10	25.0	25.0	85.0
	4.00	5	12.5	12.5	97.5
	5.00	1	2.5	2.5	100.0
Total		40	100.0	100.0	

**Fig. 6.7** Doing a frequency table in Stata


```
Command
tab Times_Partying
```

categories. Columns two, three, and five display the raw frequency, the corresponding percentage for each category, and the cumulative percentage, respectively. The fourth column, labeled valid percent, displays the percentage of each cell by taking into consideration missing values. Missing values are answers to questions left blank by the respondent (i.e., a respondent did not answer the question). In our example, all survey respondents answered the question on how many times they normally go party. Therefore, there are no missing values, and the values in column four match the values in column three.

### 6.4.2 Constructing a Frequency Table in Stata

Step 1: Write in the Stata Command field “tab Times\_Partying,” and press enter.

(Alternatively, you can also write tab and then click on the variable Times\_Partying in the upper right corner of the display.) (See Fig. 6.7.)

The Stata output in Table 6.3 has four columns: (1) variable name (this lists all the possible categories), (2) raw frequency (lists the occurrence of each category), (3) Percent per category (lists the percentage of values that fall into any category), and (4) cumulative percentage (the cumulative percentage of values that fall into the listed category or a lower category).

**Table 6.3** Stata frequency table output

Times_Party ing	Freq.	Percent	Cum.
0	3	7.50	7.50
1	8	20.00	27.50
2	10	25.00	52.50
3	9	22.50	75.00
4	8	20.00	95.00
5	2	5.00	100.00
Total	40	100.00	

## 6.5 The Measures of Central Tendency: Mean, Median, Mode, and Range

This part shortly introduces the most widely used measures of central tendency or univariate statistics: mean, median, mode, and range.

### Mean

The mean is the value we commonly call the average. To calculate the mean, sum up all observations, and then divide this sum by the number of subjects or observations.

In statistical language the mean is denoted by  $\bar{x}$  an observation is denoted  $x$ , and the sample is denoted by  $n$ .

The mean in mathematical language:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

The mean can be strongly influenced by outliers (i.e., observations that fall far from the rest of the data). To highlight, we take a subsample from our sample dataset. For example, the last five values of the variable money spent partying are 60, 60, 90, 70, and 200.

If we calculate the mean, we will get  $(60 + 60 + 90 + 70 + 200)/5 = 96$

We can clearly see that the mean is strongly influenced by the outlier 200. While the first 4 students in our subsample spend between 60 and 90 \$/week partying, the last student spends 200 dollars, which is much different from the other values. Without the outlier 200, the mean would be only 70.

### Median

The median, or “midpoint,” is the middle number of a distribution. It is less sensitive to outliers and therefore a more “resistant” measure of central tendency than the mean. To calculate the median by hand, just line all values up in order and find the middle one (or average of the two middles when  $n$ , the number of observations, is even.).

In our example, we would line up the values: 60, 60, **70**, 90, and 200, and the median would be 70.

If we were to calculate the median without the outlier 200, we would again line up the values: 60, 60, 70, and 90. We would then calculate the mean between the two middle values 60 and 70, which is 65.

### Mode

The mode is the value that occurs most often in the sample. In cases where there are several values that occur most often, the mode can consist of these several values. Taking our five values again (i.e., 60, 60, 70, 90, 200), the value that appears most often is 60, which is the mode of this subsample.

### Range

The range is a measure that provides us with some indication how widely spread our data are. It is calculated by subtracting the highest from the lowest value. In the five-value subsample used above, the range would be 140 (i.e.,  $200 - 60$ ).

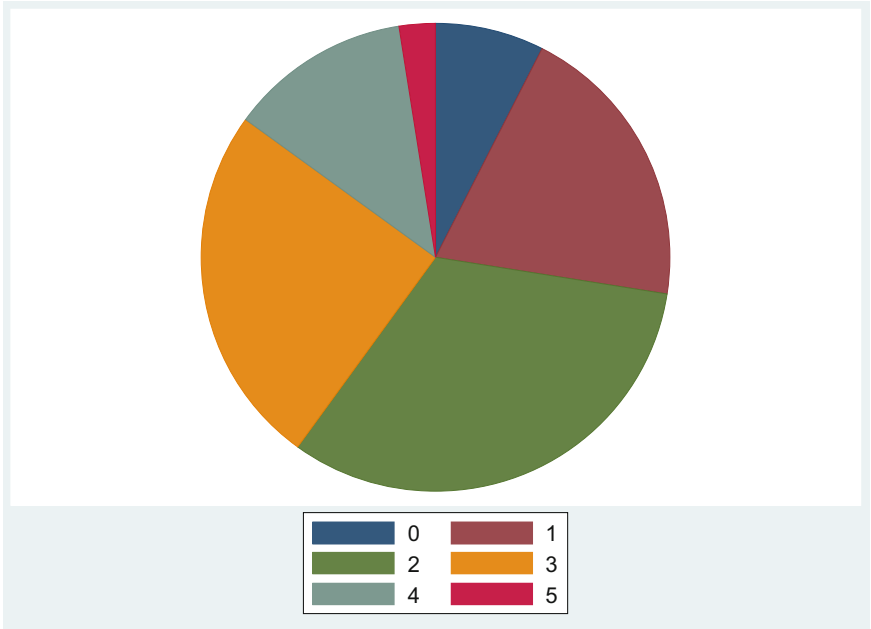
---

## 6.6 Displaying Data Graphically: Pie Charts, Boxplots, and Histograms

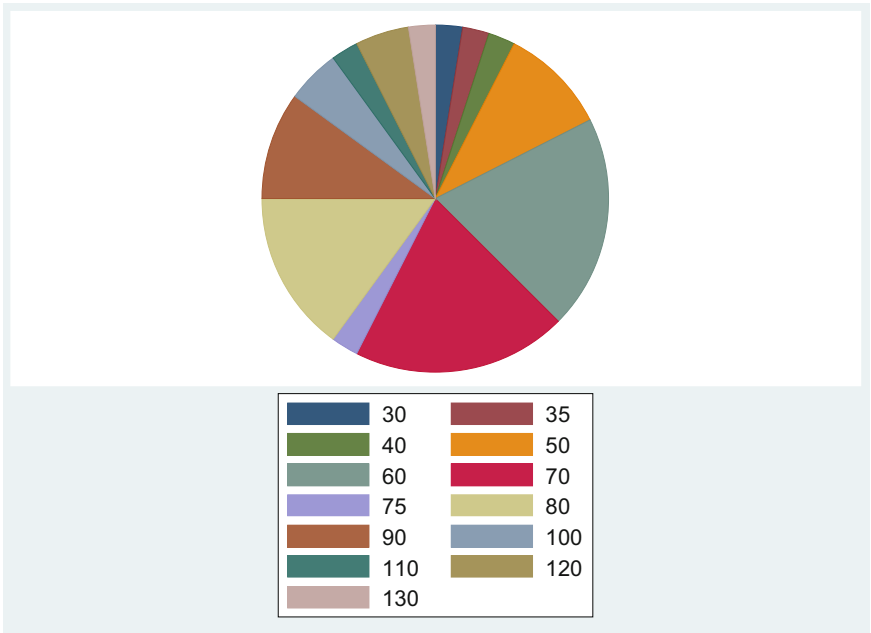
### 6.6.1 Pie Charts

A **pie chart**, sometimes also called circle chart, is one way to display a frequency table graphically. The graphic consists of a circle that is divided into slices. Each slice represents one of the variable’s categories. The size of each slice is proportional to the frequency of the value. Pie charts can be strong graphical representation of categorical data with few categories. However, the more categories we include into a pie chart the harder it is to succinctly compare categories. It is also rather difficult to compare data across different pie charts. Figure 6.8 displays the pie chart of the variable `Times_Partying`. We can see that the interpretation is already difficult, as it is already hard to guess from the graph the frequency of each slice. If we take another variable with more categories such as our dependent variable `money spent partying` (see Fig. 6.9), we see that the categories are basically indistinguishable from each other. Hence, a pie chart is not a good option to display this variable.





**Fig. 6.8** Pie chart of the variable times partying



**Fig. 6.9** Pie chart of the variable money spent partying

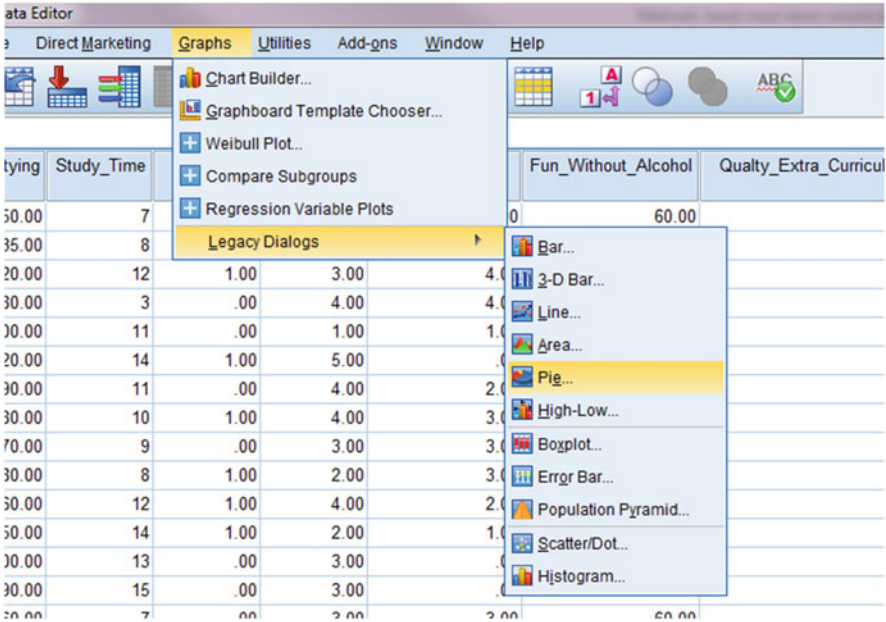


Fig. 6.10 Doing a pie chart in SPSS (first step)

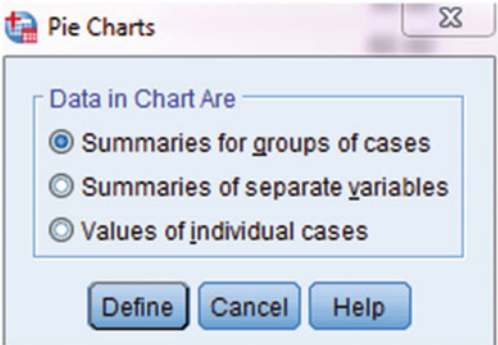
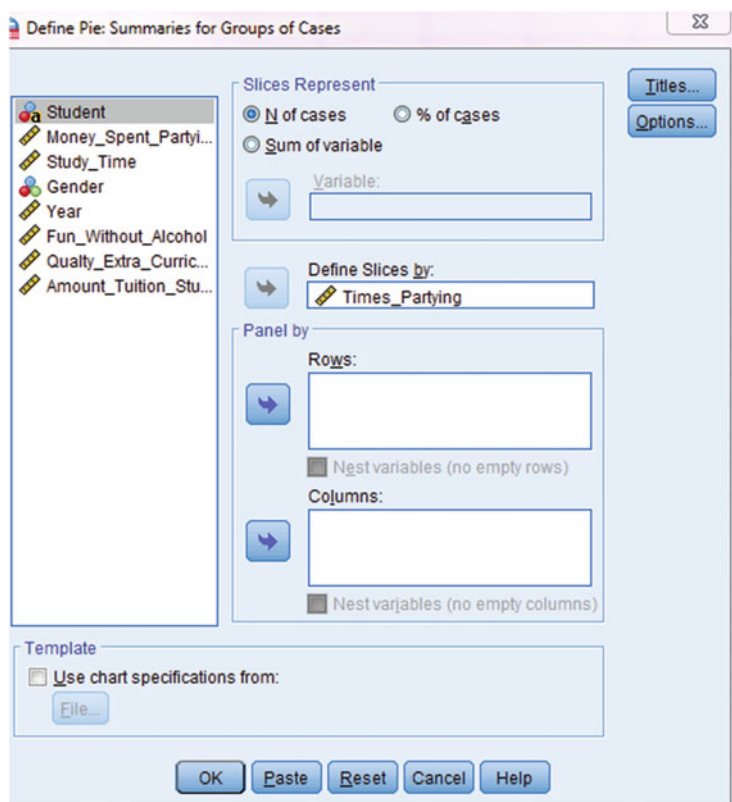


Fig. 6.11 Doing a pie chart in SPSS (second step)

6.6.2 Doing a Pie Chart in SPSS

- Step 1: Go to Graphs—Legacy Dialogue – Pie (see Fig. 6.10).
- Step 2: Click Summaries of groups of cases (see Fig. 6.11).
- Step 3: Highlight the variable—Times\_Partying—and click on the arrow next to Define Slices to include the variable in the field. Then click okay (see Fig. 6.12).



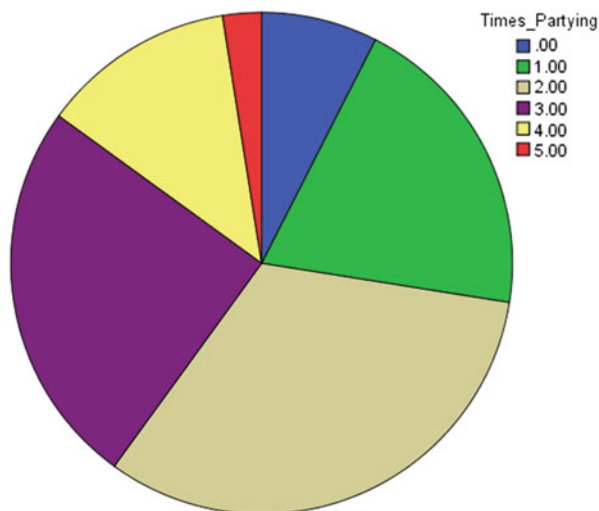
**Fig. 6.12** Doing a pie chart in SPSS (third step)

The SPSS output in Fig. 6.13 displays the pie chart; in the right-hand upper corner, it denotes the variable name and the existing categories including the corresponding colors in the chart.

### 6.6.3 Doing a Pie Chart in Stata

Step 1: Write in the Stata Command field: `graph pie, over(Times_Partying)` (see Fig. 6.14).

Figures 6.8 and 6.9 display the pie chart Stata output of the two variables times partying and money spent partying.



**Fig. 6.13** SPSS pie chart output of the variable times partying

```
Command  
graph pie, over(Times_Partying)
```

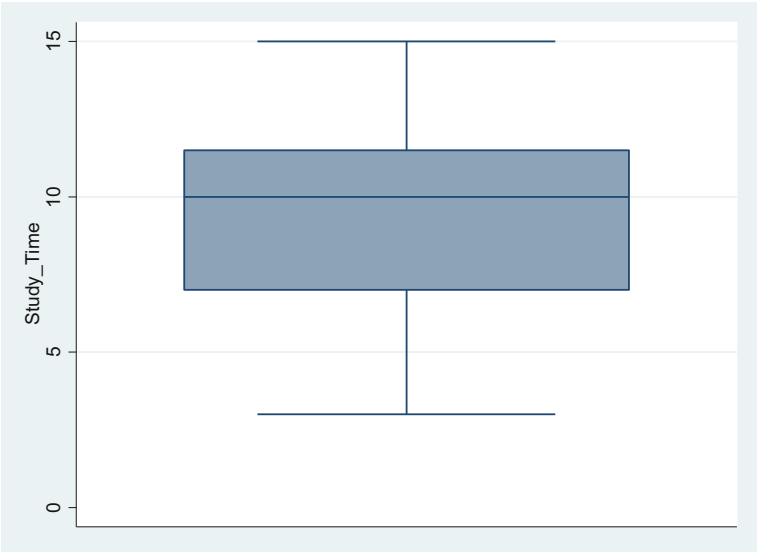
**Fig. 6.14** Doing a pie chart in Stata

## 6.7 Boxplots

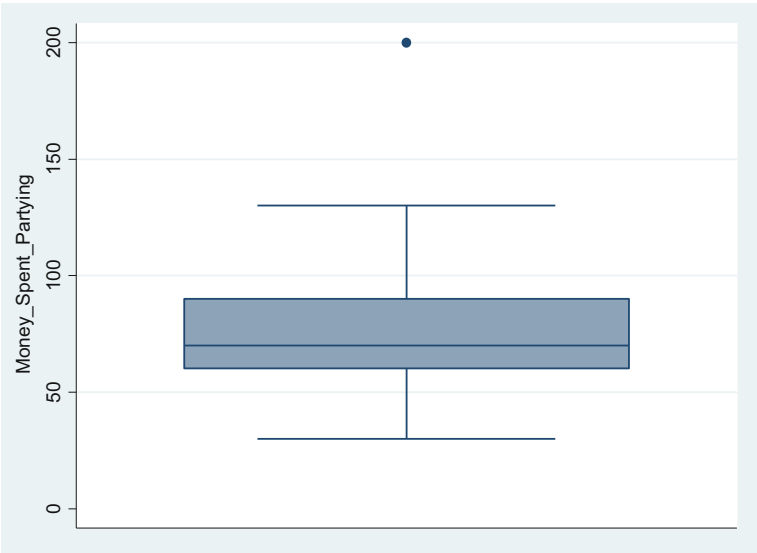
A boxplot is a very convenient way of displaying variables. This type of graph allows us to display three measures of central tendency in one graph. The median or the midpoint of each dataset is indicated by the black centerline. The blue/gray shaded box, which is also known as the interquartile range (IQR) includes the mid-50% of the data. The two outer lines denote the range of the data. If values extend up to 1.5 times the interquartile range from the upper or lower boundary of the mid-50%, they are plotted individually as asterisks. These individually plotted values are the outliers.

Figure 6.15 displays the boxplot of the variable average study time. From the graph, we see that the median study time of those students, who participated in the survey, is approximately 10 h. We also learn that the mid-50% of the data generally study between 7 and 11 h/week. The range of the data is 14. (The maximum value denoted by the upper line is 15; the minimum value denoted by the lower line is 1).

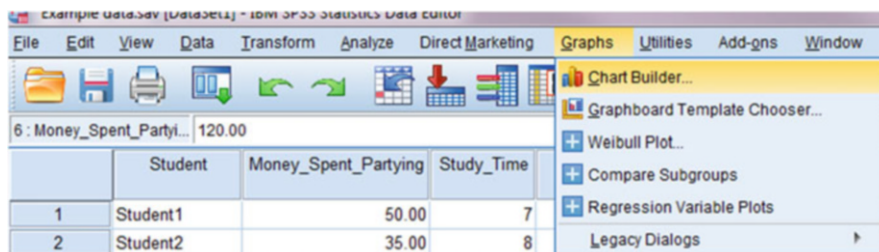
Figure 6.16 displays the boxplot of the variable—money spent partying (per week). We see that the median amount of money students spend partying is



**Fig. 6.15** Stata boxplot of the variable study time (per week)



**Fig. 6.16** Stata boxplot of the variable money spent partying (per week)



**Fig. 6.17** Doing a boxplot in SPSS (first step)

approximately 75 \$/week. The mid-50% of students spend between 60 and 90 dollars approximately. At the extremes, students spend 120 dollars at the upper end for partying and 30 dollars at the lower end. The boxplot also indicates that there is one outlying student in the data. By spending 200 dollars, she does not fit the pattern of other students.

### 6.7.1 Doing a Boxplot in SPSS

Step 1: Go to Graphs—Chart Builder; a dialogue box will open; if this is the case, press okay. You will be directed to the Chart Builder, which you see below (see Fig. 6.17).

Step 2: Go to the item—Choose from—click on Boxplot. Then in the rectangle to the right, three different types of boxplots will appear. Drag the first boxplot image to the open field above. After that, click on your variable of interest—in our case, study time—and drag it to the y-axis. Finally, click okay (see Figs. 6.18 and 6.19).

Figure 6.19 displays the boxplot of the variable study time. The median is at 10 h/week. The range is 14 h (i.e., the minimum in the sample is 1 h, the maximum is 15 h), and the interquartile range or the mid-50% of the data goes from 7 to 11.

### 6.7.2 Doing a Boxplot in Stata

Step 1: Write in the Stata Command field: `graph box Study_Time` (see Fig. 6.20).

Figures 6.15 and 6.16 display two Stata boxplot outputs featuring the variables study time per week and money spent partying per week.

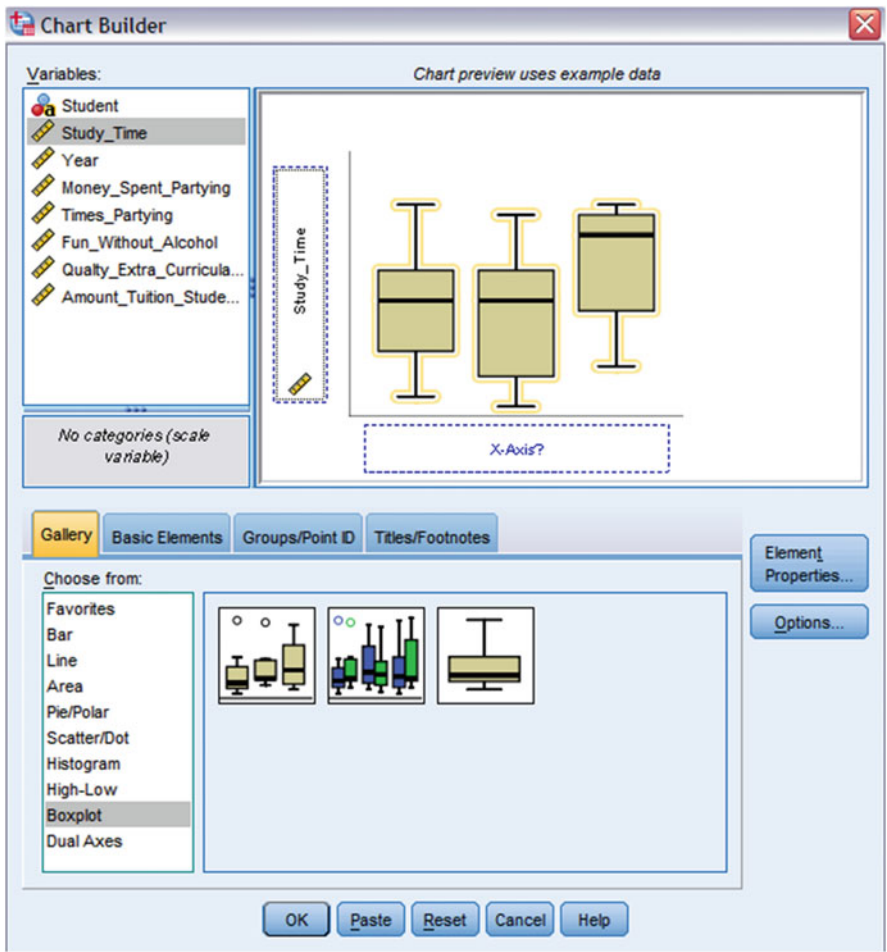


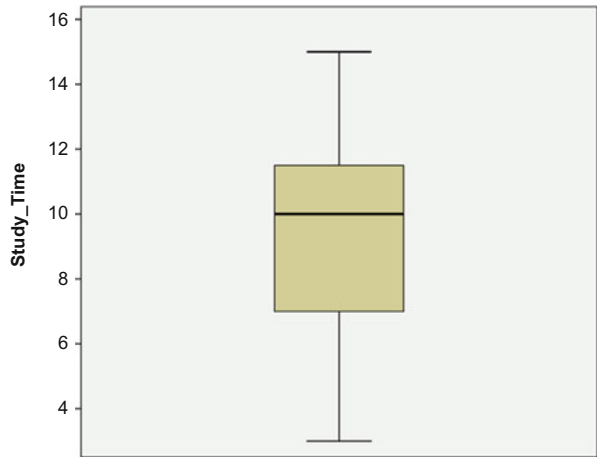
Fig. 6.18 Doing a boxplot in SPSS (second step)

## 6.8 Histograms

Histograms are one of the most widely used graphs in statistics to graphically display continuous variables. These graphs display the frequency distribution of a given variable. Histograms are very important for statistics in that they tell us if the data is normally distributed. In statistical inference—which means using a sample to generalize about a population—normally distributed data is a prerequisite for many statistical tests (see below), which we use to generalize from a sample toward a population.

Figure 6.21 shows two normal distributions (i.e., the blue line and the red line). In their ideal shape, these distributions have the following features: (1) the mode, mean,

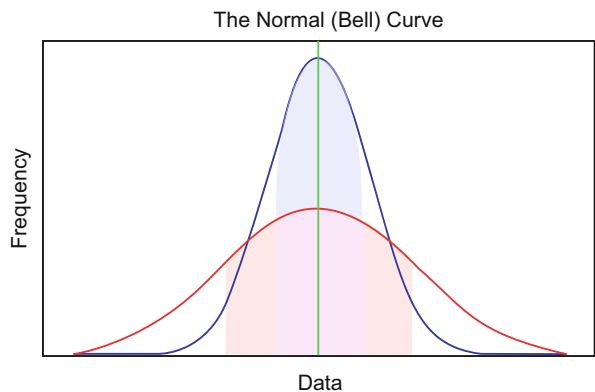
**Fig. 6.19** SPSS boxplot of the variable study time (per week)



**Fig. 6.20** Doing a boxplot in Stata

```
Command  
graph box Study_Time
```

**Fig. 6.21** Shape of a normal distribution



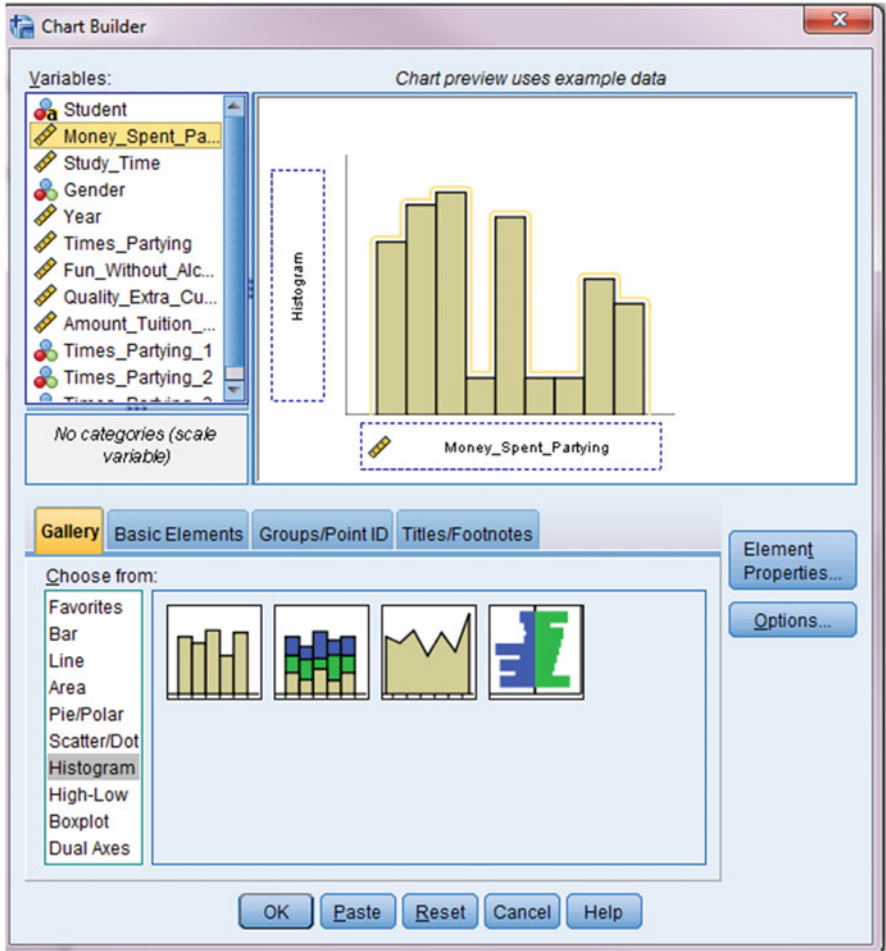
and median are the same value in the center of the distribution. (2) The distribution is symmetrical, that is, it has the same shape on each side of the distribution.

### 6.8.1 Doing a Histogram in SPSS

Step 1: Go to Graphs—Chart Builder—a dialogue box opens; when this is the case, press okay. You will be directed to the Chart Builder (this is the same procedure as constructing a boxplot).

Step 2: Go to the item “Choose from,” and click on Histogram. Then, in the rectangle to the right, four different types of histograms will appear. Drag the first type of

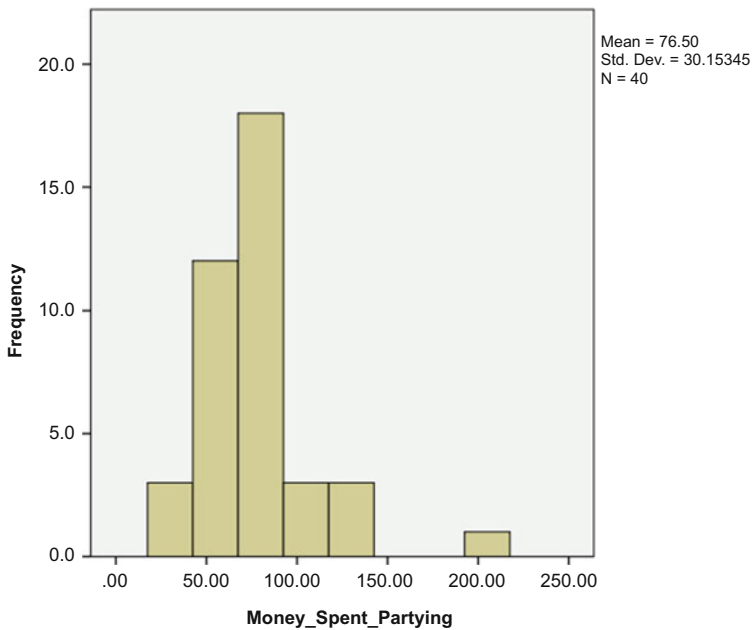




**Fig. 6.22** Doing a histogram in SPSS

histogram that appears to the open field above. After that, click on your variable of interest—in our case times spent partying—and drag it to the  $x$ -axis. Finally, click okay (see Table 7.3 and Fig. 6.22).

The histogram in Fig. 6.23 displays the distribution of the variable money spent partying. We see that the mode is approximately 80 \$/month. Pertaining to the normality assumption, we see that the data is very roughly normally distributed. There are fewer observations on the extremes and more observations in the center. However, to be perfectly normally distributed, the bar at 100 should be higher; there should also not be any outlier at 200. However, for analytical purposes, we would



**Fig. 6.23** SPSS Histogram of the variable money spent partying per week

```
Command
hist Money_Spent_Partying
```

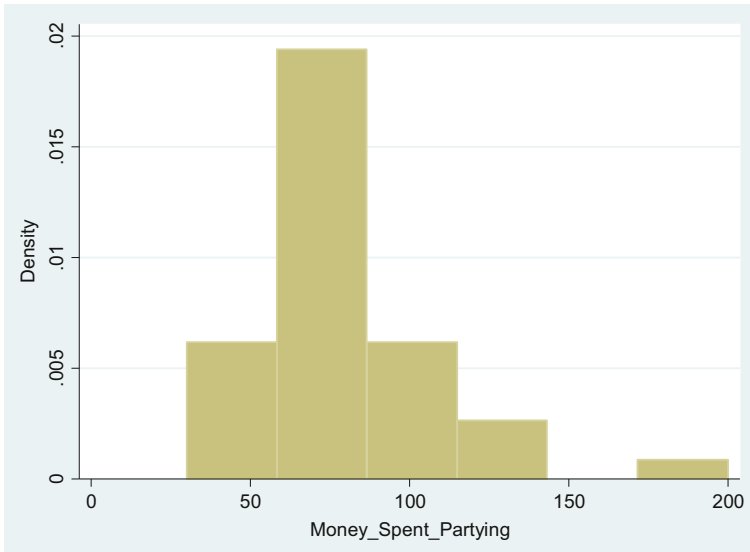
**Fig. 6.24** Doing a histogram in Stata

say that this graph is close enough to a normal distribution to make “correct” inferences from samples to populations.

## 6.8.2 Doing a Histogram in Stata

Step 1: Write in the Stata Command field: `hist Money_Spent_Partying` (see Fig. 6.24).

The Stata output of the variable money spent partying (see Fig. 6.25) uses somewhat larger bars than the SPSS output and is therefore a little less “precise” than the SPSS output.



**Fig. 6.25** Stata Histogram of the variable money spent partying (per week)

## 6.9 Deviation, Variance, Standard Deviation, Standard Error, Sampling Error, and Confidence Interval

On the following pages, I will illustrate how you can calculate the sampling error and the confidence interval, two univariate statistics that are of high value for survey research. In order to calculate the sampling error and confidence interval, we have to follow several intermediate steps. We have to calculate the deviation, sample variance, standard deviation, and standard error.

### Deviation

Every sample has a sample mean, and for each observation there is a deviation from that mean. The deviation is positive when the observation falls above the mean and negative when the observation falls below the mean. The magnitude of the value reports how different (in the relevant numerical scale) an observation is from the mean.

**Formula deviation:** Difference between the observation and the mean,  $Y_i - \hat{Y}$

Example: Assume we have the following three numbers: 1, 2, and 6

For these numbers the deviations are:

$$1 - 3 = -2$$

$$2 - 3 = -1$$

$$6 - 3 = 3$$

(By definition, the sum of these deviations is 0)

### Sample Variance

The variance is the approximate average of the squared deviations. In other words, the variance measures the approximate average of the squared distance between observations and the mean. For this measure, we use squares because the deviations can be negative, and squaring gets rid of the negative sign.

### Formula Sample Variance

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

### Standard Deviation

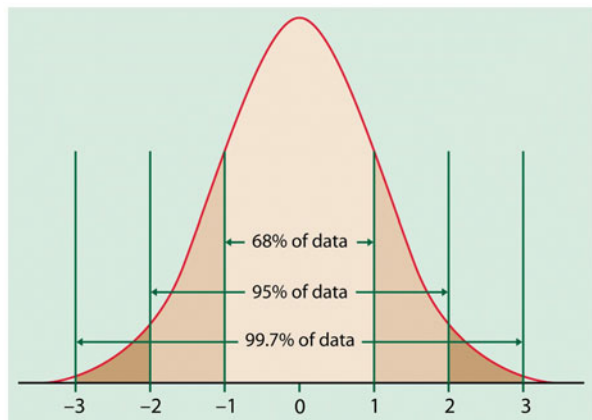
The standard deviation is a measure of volatility that measures the amount of variability or volatility around the mean. The standard deviation is large if there is high volatility in the data and low if the data is closely clustered around the mean. In other words, the smaller the standard deviation, the less “error” we have in our data and the more secure we can be in knowing that our sample mean closely matches our population mean.

### Formula Standard Deviation

$$S = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}}$$

The standard deviation is also important for standardizing variables. If the data are normally distributed (i.e., they follow a bell-shaped curve), the data have the following properties. Sixty-eight percent of the cases fall within one standard deviation, 95% of the cases fall between two standard deviations, and 99.7% cases fall between three standard deviations (see Figs. 6.26).

**Fig. 6.26** Standard deviation in a normal distribution



### Standard Error

The standard error allows researchers to measure how close the mean of the sample is to the population mean.

### Formula of the Standard Error

The diagram shows the formula for the standard error of the mean:  $\sigma_{\bar{x}} = \frac{S}{\sqrt{n}}$ . Three blue arrows point from text labels to parts of the formula:
 

- An arrow from "The standard error" points to  $\sigma_{\bar{x}}$ .
- An arrow from "The standard deviation of the sample" points to  $S$ .
- An arrow from "The sample size (n)" points to  $\sqrt{n}$ .

The standard error is important, because it allows researchers to calculate the confidence interval. The confidence interval, in turn, allows researchers to make inferences from the sample mean toward the population mean. It allows researchers to calculate the population mean based on the sample mean. In other words, it gives us a range in which the real mean falls.

(In reality, this method only works if we have a random sample and a normally distributed variable.)

### Formula of the Confidence Interval

The confidence interval applied:

The diagram shows the formula for the confidence interval:  $\bar{x} \pm z^* \left( \frac{s}{\sqrt{n}} \right)$ . Red brackets and lines connect text labels to parts of the formula:
 

- A bracket above  $\bar{x}$  is labeled "Sample Mean".
- A bracket above the entire  $\pm$  term is labeled "Margin of Error".
- A bracket below  $z^*$  points to a text box: "The z score that corresponds to how confident you want to be in your results (usually .90, .95 or .99). In statistics we want to be 95 percent confident, hence we replace z with 1.96. In a normal distribution going 1.96 standard deviations to right and left from the mean includes 95 percent of the data".
- A bracket below the fraction  $\frac{s}{\sqrt{n}}$  is labeled "The standard error of the sampling distribution".

Surveys generally use the confidence interval to depict the accuracy of their predictions.

For example, a 2006 opinion poll of 1000 randomly selected Americans aged 18–24 conducted by the Roper Public Affairs Division and National Geographic finds that:

- Sixty-three percent of young adults ages 18–24 cannot find Iraq on a map of the Middle East.
- Eighty-eight percent of young adults ages 18–24 cannot find Afghanistan on a map of Asia.

At the end of the survey, we find the stipulation that the results of this survey are accurate at the 95% confidence level  $\pm 3\%$  points (margin of error  $\pm 3\%$ ).

This means that we are 95% confident that the *true population* statistic, i.e., the *true* percentage of American youths who cannot find Iraq on a map is somewhere in between 60 and 66. In other words, the “real” mean in the population is anywhere between  $\pm 3\%$  points from the mean. This error range is normally called the margin of error or the **sampling error**. In the Iraqi example, we say we have a sampling error of  $\pm 3\%$  points (mean  $\pm 3\%$  points) (see Fig. 6.27).

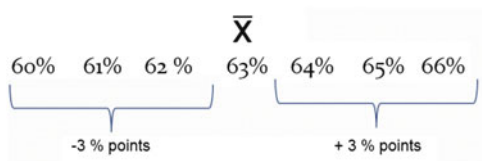
### Calculating the Confidence Interval (by Hand)

To give you an idea on how to construct the confidence interval by hand using the ten first values of the variable study time per week of our sample dataset (see Appendix 1).

Step 1: Calculating the mean

$$(7 + 8 + 12 + 3 + 11 + 14 + 11 + 10 + 9 + 8)/10 = 9.3$$

**Fig. 6.27** Graphical depiction of the confidence interval



Step 2: Calculating the variance

$$\begin{aligned} & (7 - 9.3)^2 + (8 - 9.3)^2 + (12 - 9.3)^2 + (3 - 9.3)^2 + (11 - 9.3)^2 + (14 - 9.3)^2 \\ & + (11 - 9.3)^2 + (10 - 9.3)^2 + (9 - 9.3)^2 + (8 - 9.3)^2 / 9 \\ & = 9.34 \end{aligned}$$

Step 3: Calculating the standard deviation

$$\sqrt{9.34} = 3.06$$

Step 4: Calculating the standard error

$$\left( \frac{9.34}{\sqrt{10}} \right) = 2.95$$

Step 5: Calculating the confidence interval

$$9.3 \pm 1.96 \times \left( \frac{9.34}{\sqrt{10}} \right) = 15.09; 3.51$$

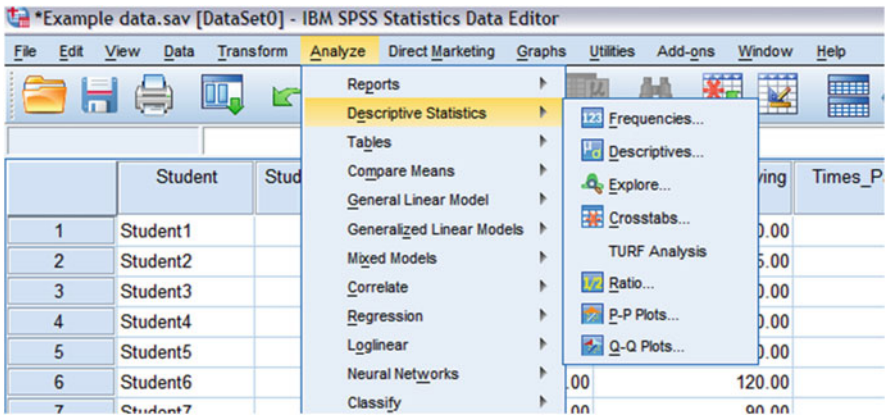
Assuming that this sample is random and normally distributed, we would find that the real average study time of students lies between 3.5 and 15.1 h. This confidence interval is large because we have few observations and relatively widespread data.

### 6.9.1 Calculating the Confidence Interval in SPSS

With the help of SPSS, we can calculate the standard deviation and the standard error. We cannot directly calculate the confidence interval but instead use the SPSS Descriptive Statistics to calculate it by hand (i.e., we have to do the last step by hand).

Step 1: Go to Analyze—Descriptive Statistics—Descriptives (see Fig. 6.28).

Step 2: Once the following window appears, drag the variable study time to the right. Then, click on options. The menu, which you will see to the right, will open. On this menu, you can choose what statistics SPSS will display. Add the option



**Fig. 6.28** Calculating the confidence interval in SPSS (first step)

S.E. mean. (The options for Mean, Std. Deviation, Minimum, and Maximum are checked automatically). Click continue and okay (see Fig. 6.29).

You will receive the following SPSS output (see Table 6.4) (please note that this output is based on data for the variable study time from the whole dataset and not only the first ten observations). The output displays the number of observations ( $N$ ), the minimum and maximum value, and the mean, accompanied by its standard error and the standard deviation. If we want to calculate the confidence interval, we have to do it by hand by using the formula introduced above.

The confidence interval for the variable study time is:

Calculating the upper limit:  $9.38 + 1.96 \times 0.489$

Calculating the lower limit:  $9.38 - 1.96 \times 0.489$

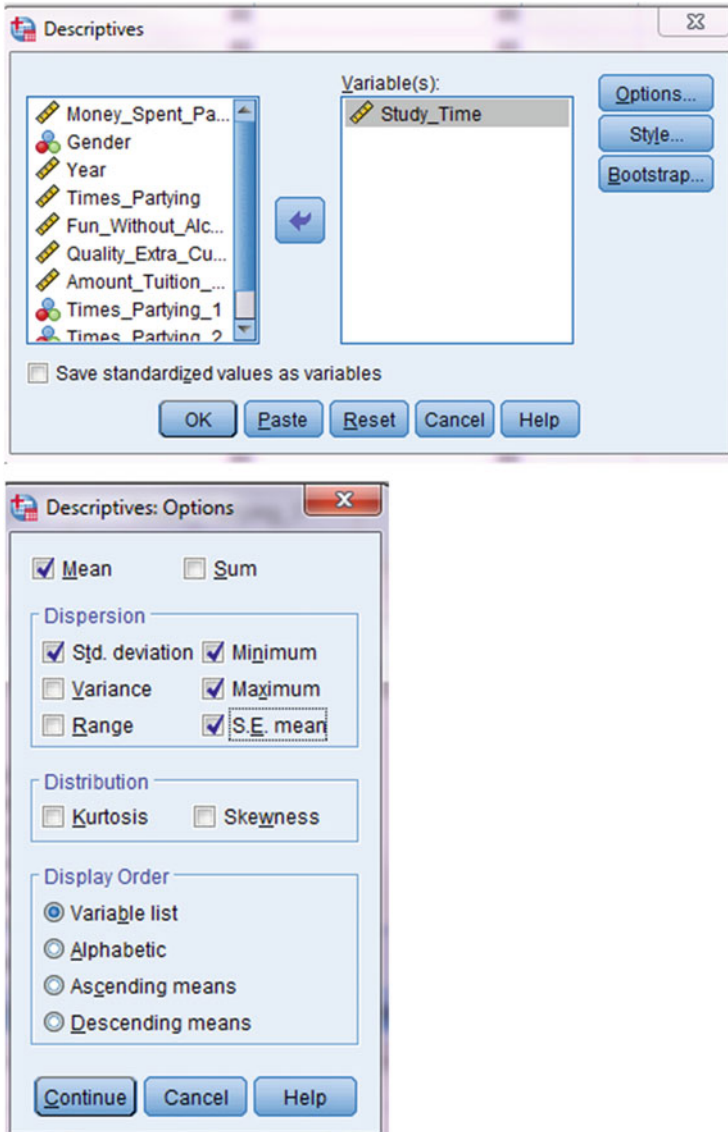
Assuming that the questionnaire data (see appendix) was drawn from a random sample of students that is normally distributed, we could conclude with 95% certainty that the real mean in students' study time would lie between 8.42 and 10.34 h (see Table 6.4).

### 6.9.2 Calculating the Confidence Interval in Stata

Step 1: Write in the Stata Command field: `tabstat Study_Time, stats(mean sd semean min max n)` (see Fig. 6.30).

(mean = mean; sd = standard deviation; semean = standard error of the mean; min = minimum; max = maximum; max = maximum;  $n$  = number of observations)





**Fig. 6.29** Calculating the confidence interval in SPSS (second step)

The stat output provides five statistics: (1) the number of observation the calculations are based on, (2) the sample mean, (3) the standard deviation, (4) the minimum sample value, and (5) the maximum sample value. The confidence interval is not explicitly listed. If we want to calculate it, we can do so by hand (see also Table 6.5):

**Table 6.4** Descriptive statistics of the variable study time (SPSS output)

Descriptive statistics						
	<i>N</i>	Minimum	Maximum	Mean		Std. deviation
	Statistic	Statistic	Statistic	Statistic	Std. error	Statistic
Study_Time	40	3	15	9.38	0.489	3.094
Valid <i>N</i> (listwise)	40					

```
Command
tabstat Study_Time, stats(mean sd semean min max n)
```

**Fig. 6.30** Calculating the confidence interval in Stata**Table 6.5** Descriptive statistics of the variable study time (Stata output)

variable	mean	sd	se(mean)	min	max	N
Study_Time	9.375	3.094143	.4892269	3	15	40

Calculating the upper limit:  $9.38 + 1.96 \times 0.489$

Calculating the lower limit:  $9.38 - 1.96 \times 0.489$

Assuming that the questionnaire data (see appendix) would be drawn from a random sample of students that is normally distributed, we could conclude with 95% certainty that the real mean in students' study time would lie between 8.42 and 10.34 h.

## Further Reading

### SPSS Introductory Books

- Cronk, B. C. (2017). *How to use SPSS®: A step-by-step guide to analysis and interpretation*. London: Routledge. Hands-on introduction into the statistical package SPSS designed for beginners. Shows users how to enter data and conduct some rather simple statistical tests.
- Green, S. B., & Salkind, N. J. (2016). *Using SPSS for Windows and Macintosh, Books a la Carte*. Upper Saddle River: Pearson. An introduction into SPSS specifically designed for students of the social and political sciences. Guides users through basic SPSS techniques and statistics.

## Stata Introductory Books

Mehmetoglu, M., & Jakobsen, T. G. (2016). *Applied statistics using Stata: A guide for the social sciences*. London: Sage. A good applied textbook into regression analysis with plenty of applied examples in Stata.

Pollock III, P. H. (2014). *A Stata® companion to political analysis*. Thousand Oaks: CQ Press. Provides a step-by-step introduction into Stata. It includes plenty of supplementary material such as a sample dataset, more than 50 exercises and customized screenshots.

## Univariate and Descriptive Statistics

Park, H. M. (2008). *Univariate analysis and normality test using SAS, Stata, and SPSS*. Technical working paper. The University Information Technology Services (UITs) Center for Statistical and Mathematical Computing, Indiana University. <https://scholarworks.iu.edu/dspace/handle/2022/19742>. A concise introduction into descriptive statistics and graphical representations of data including a discussion of their underlying statistical assumptions.