Chapter 2
Exploratory and Descriptive Methods

Before beginning any sort of statistical analysis, it is imperative to take a preliminary look at the data with three main goals in mind: first, to check for errors and anomalies; second, to understand the distribution of each of the variables on its own; and third, to begin to understand the nature and strength of relationships among variables. Errors should, of course, be corrected, since even a small percentage of erroneous data values can drastically influence the results. Understanding the distribution of the variables, especially the outcomes, is crucial to choosing the appropriate multipredictor regression method. Finally, understanding the nature and strength of relationships is the first step in building a more formal statistical model from which to draw conclusions.

2.1 Data Checking

Procedures for data checking should be implemented before data entry begins, to head off future headaches. Many data entry programs have the capability to screen for egregious errors, including values that are out the expected range or of the wrong "type." If this is not possible, then we recommend regular checking for data problems as the database is constructed.

Here are two examples we have encountered recently. First, some values of a variable defined as a proportion were inadvertently entered as percentages (i.e., 100 times larger than they should have been). Although they made up less than 3% of the values, the analysis was completely invalidated. Fortunately, this simple error was easily corrected once discovered. A second example involved patients with a heart anomaly. Those whose diagnostic score was poor enough (i.e., exceeded a numerical threshold) were to be classified according to the type of anomaly. Data checks revealed missing classifications for patients whose diagnostic score exceeded the

E. Vittinghoff et al., *Regression Methods in Biostatistics*, Statistics for Biology and Health, DOI 10.1007/978-1-4614-1353-0_2,

threshold, as well as classifications for patients whose score did not, complicating planned analyses. Had the data been screened as they were collected, this problem with study procedures could have been avoided.

2.2 Types of Data

The proper description of data depends on the nature of the measurement. The key distinction for statistical analysis is between numerical and categorical variables. The number of diagnostic tests ordered is a numerical variable, while the gender of a person is categorical. Systolic blood pressure (SBP) is numerical, whereas the type of surgery is categorical.

A secondary but sometimes important distinction within numerical variables is whether the variable can take on a whole continuum or just a discrete set of values. So SBP would be continuous, while number of diagnostic tests ordered would be discrete. Cost of a hospitalization would be continuous, whereas number of mice able to successfully navigate a maze would be discrete. More generally,

> *Definition*: A numerical variable taking on a continuum of values is called *continuous* and one that only takes on a discrete set of values is called *discrete*.

A secondary distinction sometimes made with regard to categorical variables is whether the categories are ordered or unordered. So, for example, categories of annual household income (<$20,000, $20,000–$40,000, $40,000–$100,000, >$100,000) would be ordered, while marital status (single, married, divorced, widowed) would be unordered. More exactly,

> *Definition*: A categorical variable is *ordinal* if the categories can be logically ordered from smallest to largest in a sense meaningful for the question at hand (we need to rule out silly orders like alphabetical); otherwise it is unordered or *nominal*.

Some overlap between types is possible. For example, we may break a numerical variable (such as exact annual income in dollars and cents) into ranges or categories. Conversely, we may treat a categorical variable as a numerical score, for example, by assigning values one to five to the ordinal responses Poor, Fair, Good, Very Good, and Excellent.

Most of the analysis methods we will describe for numerical scores (e.g., linear regression or t-tests) have interpretations based on average scores. So assigning scores to a categorical variable is effective if average scores are readily interpretable. This may well be the case for scoring the categories Poor through Excellent as 1 through 5: an average value of 3.5 is between Good and Very Good. It might be a less effective strategy ordinal categorical variables such as the modified Rankin Scale, a scale used to assess disability following a stroke. For that scale, 0 represents no symptoms, 1 and 2 slight disability, 3 and 4 moderate disability, 5 severe disability, and 6 is dead. Consider two sets of three patients, the first set with scores of 0, 6, and 6 and the second with scores of 4, 4, and 4. Both have averages of 4, but the

first set would generally be considered as having worse outcomes since two of the patients died. In such a case, summarizing with the average, and hence treating the variable as numeric, may not be appropriate.

In the following sections, we present each of the descriptive and exploratory methods according to the types of variables involved.

2.3 One-Variable Descriptions

We begin by describing techniques useful for examining a single variable at a time. These are useful for uncovering mistakes or extreme values in the data and for assessing distributional shape.

2.3.1 Numerical Variables

We can describe the distribution of numerical variables using either numerical or graphical techniques.

2.3.1.1 Example: Systolic Blood Pressure

The western collaborative group study (WCGS) was a large epidemiological study designed to investigate the association between the "type A" behavior pattern and coronary heart disease (CHD) (Rosenman et al. 1964). We will revisit this study later in the book, focusing on the primary outcome, but for now we want to explore the distribution of SBP.

2.3.1.2 Numerical Description

As a first step, we obtain basic descriptive statistics for SBP. Table 2.1 gives detailed summary statistics for the SBP variable, `sbp`. Several features of the output are worth consideration. The largest and smallest values should be scanned for outlying or incorrect values, and the mean (or median) and standard deviation should be assessed as general measures of the location and spread of the data. Secondary features are the skewness and kurtosis, though these are usually more easily assessed by the graphical means described in the next section. Another assessment of skewness is a large difference between the mean and median. In *right-skewed* data, the mean is quite a bit larger than the median, while in *left-skewed* data, the mean is much smaller than the median. Of note, in this dataset, the largest observation is more than six standard deviations above the mean!

Table 2.1 Numerical description of systolic blood pressure

```
. summarize sbp, detail

                        systolic BP
-------------------------------------------------------------
      Percentiles      Smallest
 1%           104            98
 5%           110           100
10%           112           100       Obs                3154
25%           120           100       Sum of Wgt.        3154

50%           126                     Mean           128.6328
                        Largest       Std. Dev.      15.11773
75%           136           210
90%           148           210       Variance       228.5458
95%           156           212       Skewness       1.204397
99%           176           230       Kurtosis       5.792465
```

2.3.1.3 Graphical Description

Graphs are often the quickest and most effective way to get a sense of the data. For numerical data, three basic graphs are most useful: the histogram, boxplot, and normal quantile–quantile (or Q–Q) plot. Each is useful for different purposes. The histogram easily conveys information about the location, spread, and shape of the frequency distribution of the data. The boxplot is a schematic identifying key features of the distribution. Finally, the normal Q–Q plot facilitates comparison of the shape of the distribution of the data to a normal (or bell-shaped) distribution.

The histogram displays the frequency of data points falling into various ranges as a bar chart. Figure 2.1 shows a histogram of the SBP data from WCGS. Generated using an earlier version of Stata, the default histogram uses five intervals and labels axes with the minimum and maximum values only. In this figure, we can see that most of the measurements are in the range of about 100 to 150, with a few extreme values around 200. The percentage of observations in the first interval is about 47.4%.

However, this is not a particularly well-constructed histogram. With over 3,000 data points, we can use more intervals to increase the definition of the histogram and avoid grouping the data so coarsely. Using only five intervals, the first two including almost all the data, makes for a loss of information, since we only know the value of the data in those large "bins" to the limits of the interval (in the case of the first bin, between 98 and 125), and learn nothing about how the data are distributed within those intervals. Also, our preference is to provide more interpretable axis labeling. Figure 2.2 shows a modified histogram generated using the current version of Stata that provides much better definition as to the shape of the frequency distribution of SBP.

The key with a histogram is to use a sufficient number of intervals to define the shape of the distribution clearly and not lose much information, without using so many as to leave gaps, give the histogram a ragged shape, and defeat the goal of summarization. With 3,000 data points, we can afford quite a few bins. A *rough*

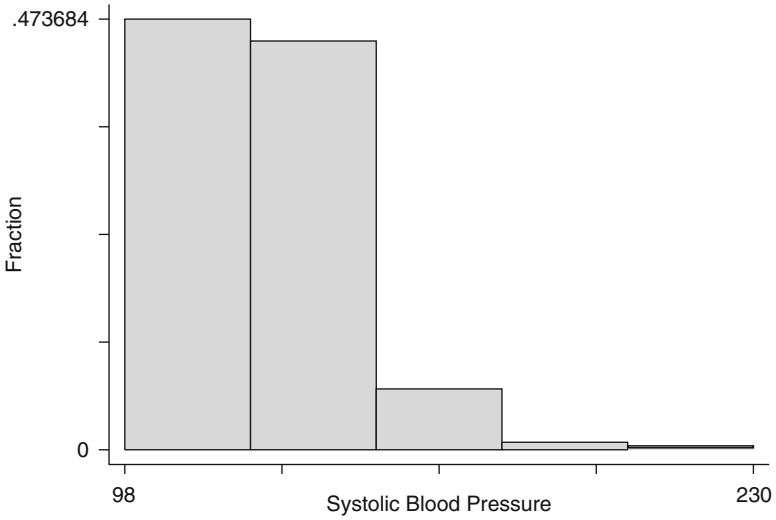

Fig. 2.1 Histogram of the systolic blood pressure data

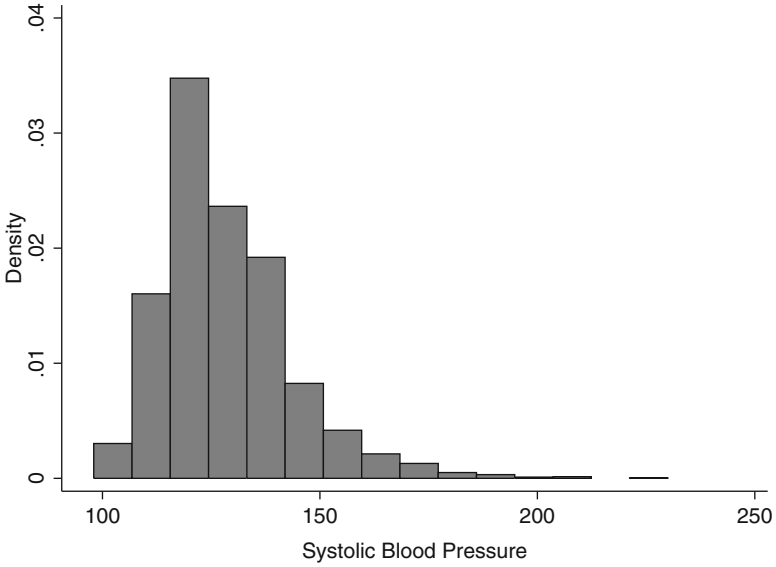

Fig. 2.2 Histogram of the systolic blood pressure data using 15 intervals

rule of thumb is to choose the number of bins to be about $1 + 3.3 \log_{10}(n)$, (Sturges 1926) where n is the sample size (so this would suggest 12 or 13 bins for the WCGS data). More than 20 or so are rarely needed. Figure 2.2 uses 15 bins and provides a clear definition of the shape as well as a fair bit of detail.

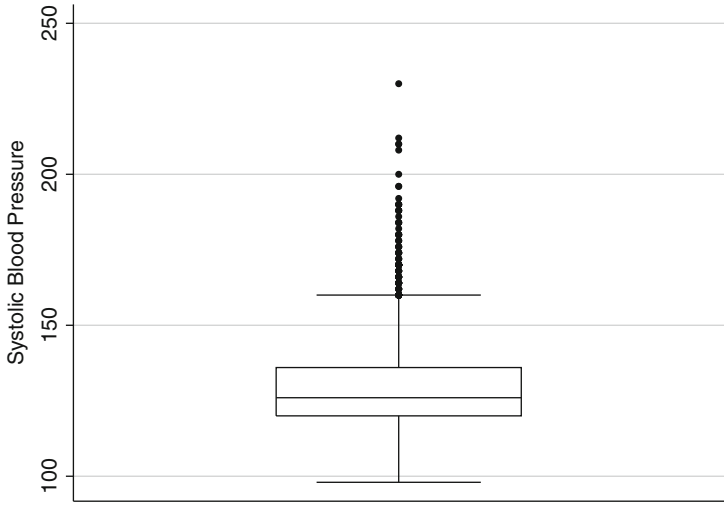

Fig. 2.3 Boxplot of the systolic blood pressure data

A boxplot represents a compromise between a histogram and a numerical summary. The boxplot in Fig. 2.3 graphically displays information from the summary in Table 2.1, specifically the minimum, maximum, and 25th, 50th (median), and 75th percentiles. This retains many of the advantages of a graphical display while still providing fairly precise numerical summaries. The "box" displays the 25th and 75th percentiles (the lower and upper edges of the box) and the median (the line across the middle of the box). Extending from the box are the "whiskers" (this colorful terminology is due to the legendary statistician John Tukey, who liked to coin new terms). The bottom whisker extends to the minimum data value, 98, but the maximum is above the upper whisker. This is because Stata uses an algorithm to try to determine if observations are "outliers," that is, values a large distance away from the main portion of the data. Data points considered outliers (they can be in either the upper or lower range of the data) are plotted with symbols and the whisker only extends to the most extreme observation not considered an outlier.

Boxplots convey a wealth of information about the distribution of the variable:

- Location, as measured by the median
- Spread, as measured by the height of the box (this is called the interquartile range or IQR)
- Range of the observations
- Presence of outliers
- Some information about shape

This last point bears further explanation. If the median is located toward the bottom of the box, then the data are *right-skewed* toward larger values. That is, the

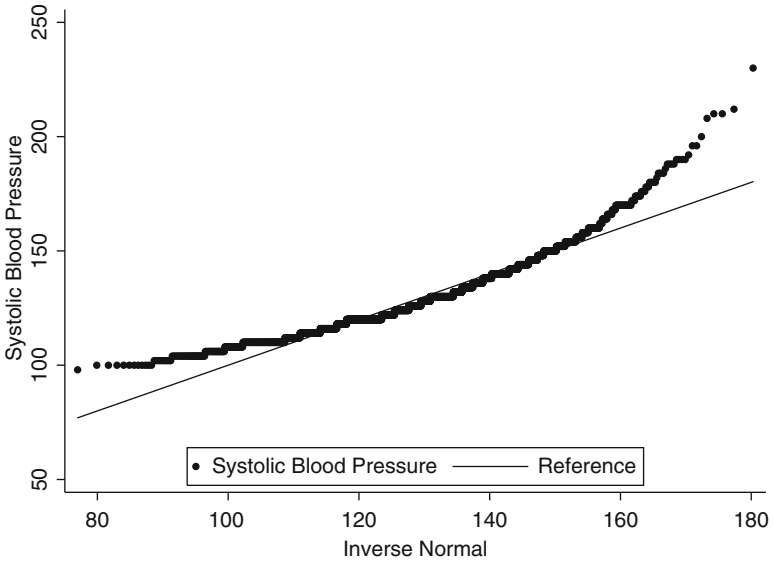

Fig. 2.4 Normal Q–Q plot of the systolic blood pressure data

distance between the median and the 75th percentile is greater than that between the median and the 25th percentile. Likewise, right-skewness will be indicated if the upper whisker is longer than the lower whisker or if there are more outliers in the upper range. Both the boxplot and the histogram show evidence for right-skewness in the SBP data. If the direction of the inequality is reversed (more outliers on the lower end, longer lower whisker, median toward the top of the box), then the distribution is *left-skewed*.

Our final graphical technique, the normal Q–Q plot, is useful for comparing the frequency distribution of the data to a normal distribution. Since it is easy to distinguish lines that are straight from ones that are not, a normal Q–Q plot is constructed so that the data points fall along an approximately straight line when the data are from a normal distribution, and deviate *systematically* from a straight line when the data are from other distributions. Figure 2.4 shows the Q–Q plot for the SBP data. The line of the data points shows a distinct curvature, indicating the data are from a nonnormal distribution.

The shape and direction of the curvature can be used to diagnose the deviation from normality. Upward curvature, as in Fig. 2.4, is indicative of right-skewness, while downward curvature is indicative of left-skewness. The other two common patterns are S-shaped. An S-shape as in Fig. 2.5 indicates a *heavy-tailed* distribution, while an S-shape like that in Fig. 2.6 is indicative of a *light-tailed* distribution.

Heavy- and light-tailed are always in reference to a hypothetical normal distribution with the same spread. A heavy-tailed distribution has more observations in

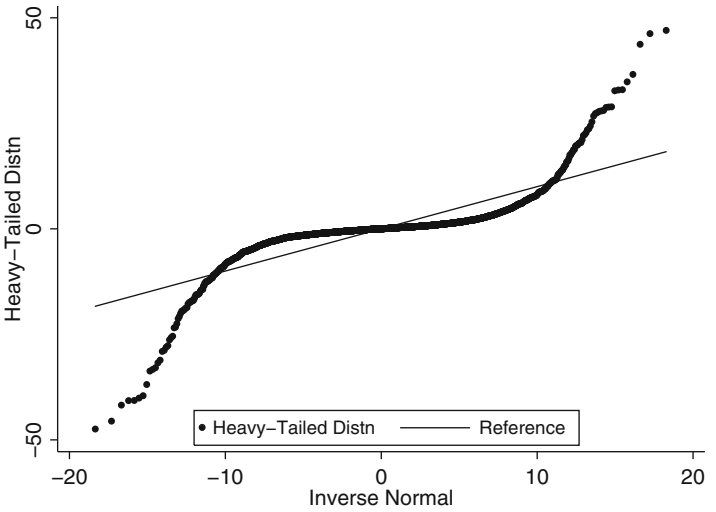

Fig. 2.5 Normal Q–Q plot of data from a heavy-tailed distribution

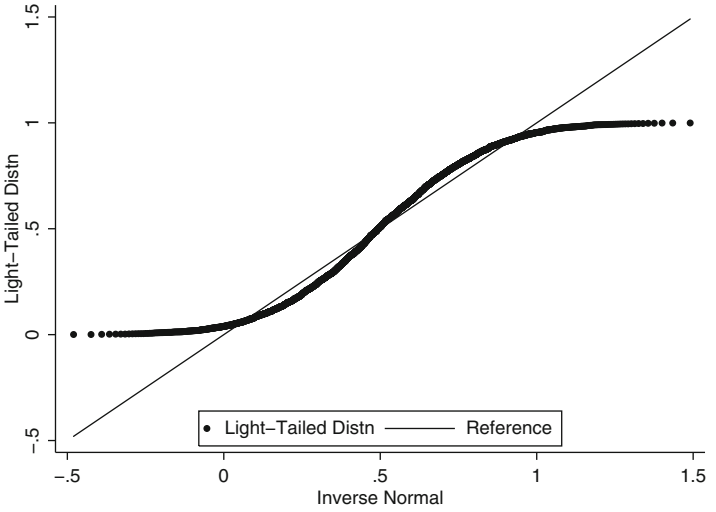

Fig. 2.6 Normal Q–Q plot of data from a light-tailed distribution

the middle of the distribution and way out in the tails, and fewer a modest way from the middle (simply having more in the tails would just mean a larger spread). Light-tailed means the reverse: fewer in the middle and far out tails and more in the mid-range. Heavy-tailed distributions are generally more worrisome than light-tailed since they are more likely to include outliers.

Table 2.2 Effect of a $\log_{10}$ transformation

Value	Difference	$\log_{10}$ value	Difference
0.01	0.09	−2	1
0.1	0.9	−1	1
1	9	0	1
10	90	1	1
100	900	2	1
1,000	–	3	–

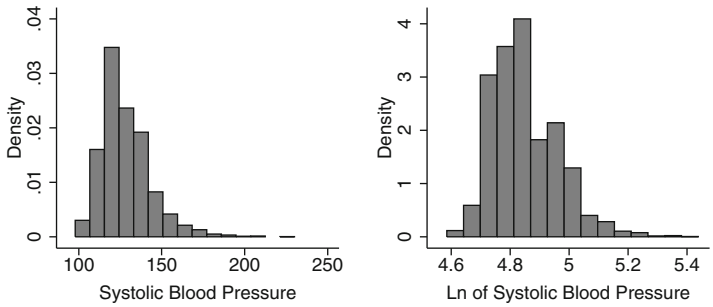

Fig. 2.7 Histograms of systolic blood pressure and its natural logarithm

2.3.1.4 Transformations of Data

A number of the techniques we describe in this book require the assumption of approximate normality or, at least, work better when the data are not highly skewed or heavy-tailed, and do not include extreme outliers. A common method for dealing with these problems is to transform such variables. For example, instead of the measured values of SBP, we might instead use the logarithm of SBP. We first consider why this works and then some of the advantages and disadvantages of transformations.

Transformations affect the distribution of values of a variable because they emphasize differences in a certain range of the data, while de-emphasizing differences in others. Consider a table of transformed values, as displayed in Table 2.2. On the original scale the difference between 0.01 and 0.1 is 0.09, but on the $\log_{10}$ scale, the difference is 1. In contrast, the difference between 100 and 1,000 on the original scale is 900, but this difference is also 1 on the $\log_{10}$ scale. So a log transformation de-emphasizes differences at the upper end of the scale and emphasizes those at the lower end. This holds for the natural log as well as $\log_{10}$ transformation. The effect can readily be seen in Fig. 2.7, which displays histograms of SBP on the original scale and after natural log transformation.

The log-transformed data is distinctly less right-skewed, even though some skewness is still evident. Essentially, we are viewing the data on a different scale of measurement.

There are a couple of other reasons to consider transforming variables, as we will see in later sections and chapters: transformations can simplify the relationships

Table 2.3 Frequencies of behavior patterns

```
tabulate behpat
 behavioral |
 pattern (4 |
     level) |      Freq.     Percent        Cum.
------------+-----------------------------------
         A1 |        264        8.37        8.37
         A2 |       1325       42.01       50.38
         B3 |       1216       38.55       88.93
         B4 |        349       11.07      100.00
------------+-----------------------------------
      Total |       3154      100.00
```

between variables (e.g., by making a curvilinear relationship linear), can remove interactions, and can equalize variances across subgroups that previously had unequal variances.

A primary objection to the use of transformations is that they make the data less interpretable. After all, who thinks about medical costs in log dollars? In situations where there is good reason to stay with the original scale of measurement (e.g., dollars), we may prefer alternatives to transformation including GLMs and weighted analyses. Or we may appeal to the robustness of normality-based techniques: many perform extremely well even when used with data exhibiting fairly serious violations of the assumptions.

In other situations, with a bit of work, it is straightforward to express the results on the original scale when the analysis has been conducted on a transformed scale. For example, Sect. 4.7.5 gives the details for log transformations in linear regression.

A compromise when the goal is, for example, to test for differences between two arms in a clinical trial is to plan ahead to present basic descriptive statistics in the original scale, but perform tests on a transformed scale more appropriate for statistical analysis. After all, a difference on the transformed scale is still a difference between the two arms.

Finally, we remind the reader that different scales of measurement just take a bit of getting used to: consider pH.

2.3.2 Categorical Variables

Categorical variables require a different approach, since they are less amenable to graphical analyses and because common statistical summaries, such as mean and standard deviation, are inapplicable. Instead we use tabular descriptions. Table 2.3 gives the frequencies, percents, and cumulative percents for each of the behavior pattern categories for the WCGS data. Note that cumulative percentages are really only useful with ordinal categorical data (why?).

When tables are generated by the computer, there is usually little latitude in the details. However, when tables are constructed by hand, thought should be given to their layout; Ehrenberg (1981) is recommended reading. Three easy-to-follow

Table 2.4 Characteristics of top medical schools

School	Rank	NIH research ($10 millions)	Tuition ($ thousands)	Average MCAT
Harvard	1	68	30	11.1
Johns Hopkins	2	31	29	11.2
Duke	3	16	31	11.6
Penn	4(Tie)	33	32	11.7
Washington U.	4(Tie)	25	33	12.0
Columbia	6	24	33	11.7
UCSF	7	24	20	11.4
Yale	8	22	30	11.1
Stanford	9(Tie)	19	30	11.1
Michigan	9(Tie)	20	29	11.0

Source: US News and World Report (http://www.usnews.com, 12/6/01)

suggestions from that article are to arrange the categories in a meaningful way (e.g., not alphabetically), report numbers to two effective digits, and to leave a gap every three or four rows to make it easier to read across the table. Table 2.4 illustrates these concepts. With the table arranged in order of the rankings, it is easy to see values that do not follow the pattern predicted by rank, for example, out-of-state tuition.

2.4 Two-Variable Descriptions

Most of the rest of this book is about the relationships among variables. An example from the WCGS is whether behavior pattern is related to SBP. In investigating the relationships between variables, it is often useful to distinguish the role that the variables play in an analysis.

2.4.1 Outcome Versus Predictor Variables

A key distinction is whether a variable is being predicted by the remaining variables, or whether it is being used to make the prediction. The variable singled out to be predicted from the remaining variables we will call the *outcome variable*; alternate and interchangeable names are *response variable* or *dependent variable*. The variables used to make the prediction will be called *predictor variables*. Alternate and equivalent terms are *covariates* and *independent variables*. We slightly prefer the outcome/predictor combination, since the term *response* conveys a cause-and-effect interpretation, which may be inappropriate, and *dependent/independent* is confusing with regard to the notion of statistical independence. ("Independent variables do not have to be independent" is a true statement!).

Table 2.5 Correlation coefficient for systolic blood pressure and weight

```
. correlate sbp weight (obs=3154)

             |      sbp   weight
-------------+------------------
         sbp |   1.0000
      weight |   0.2532   1.0000
```

In the WCGS example, we might hypothesize that change in behavior pattern (which is potentially modifiable) might cause change in SBP. This would lead us to consider SBP as the outcome and behavior pattern as the predictor.

2.4.2 Continuous Outcome Variable

As before, it is useful to consider the nature of the outcome and predictor variables in order to choose the appropriate descriptive technique. We begin with continuous outcome variables, first with a continuous predictor and then with a categorical predictor.

2.4.2.1 Continuous Predictor

When both the predictor and outcome variables are continuous, the typical numerical description is a correlation coefficient and its graphical counterpart is a scatterplot. Again considering the WCGS study, we will investigate the relationship between SBP and weight.

Table 2.5 shows the Stata command and output for the correlation coefficient, while Fig. 2.8 shows a scatterplot. Both the graph and the numerical summary confirm the same thing: there is a weak association between the two variables, as measured by the correlation of 0.25. The graph conveys important additional information. In particular, there are quite a few outliers, including an especially anomalous data point with high blood pressure and the lowest weight in the dataset.

The Pearson correlation coefficient r, more fully described in Sect. 3.2, is a scale-free measure of association that does not depend on the units in which either SBP or weight is measured. The correlation coefficient varies between –1 and 1, and correlations of absolute value 0.7 or larger are considered strong associations in many contexts. In fields where data are typically noisy, including our SBP example, much smaller correlations may be considered meaningful.

It is important to keep in mind that the Pearson correlation coefficient only measures the strength of the *linear* relationship between two variables. To determine whether the correlation coefficient is a reasonable numerical summary of the association, a graphical tool that helps to assess linearity in the scatterplot is a *scatterplot smoother.* Figure 2.9 shows a scatterplot smooth superimposed on the

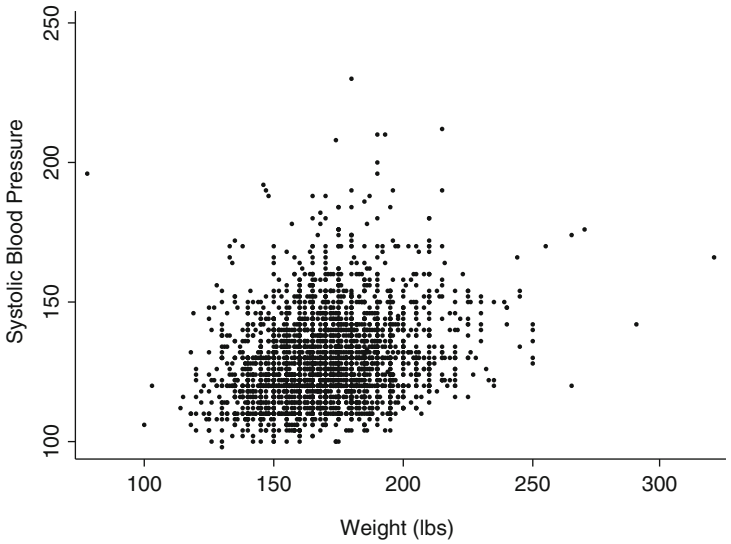

Fig. 2.8 Scatterplot of systolic blood pressure versus weight

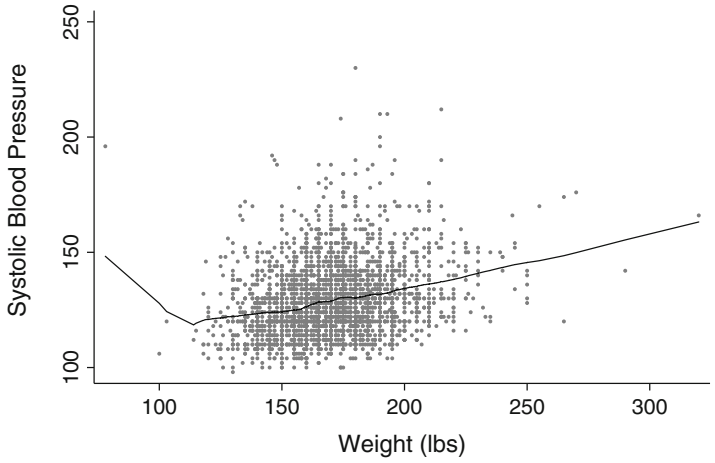

Fig. 2.9 LOWESS smooth of systolic blood pressure versus weight

graph of SBP versus weight. The figure was generated by the Stata command `lowess sbp weight, bw(0.25)` (with a few embellishments to make it look nicer). This uses the LOWESS technique to draw a smooth (but not necessarily straight) line representing the average value of the variable on the y-axis as a function of the variable on the x-axis. LOWESS is short for LOcally WEighted Scatterplot Smoother. The `bw(0.25)` option specifies that for estimation of the height of the curve at each point, 25% of the data nearest that point should be used. This is all just a fancy way of drawing a flexible curve through a cloud of points.

Table 2.6 Summary data for systolic blood pressure by behavior pattern

```
. bysort behpat: summarize sbp

-------------------------------------------------------------------------------
-> behpat = A1

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |     264    129.2462    15.29221        100        200

-------------------------------------------------------------------------------
-> behpat = A2

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |    1325    129.8891    15.77085        100        212

-------------------------------------------------------------------------------
-> behpat = B3

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |    1216    127.5551    14.78795         98        230

-------------------------------------------------------------------------------
-> behpat = B4

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         sbp |     349    127.1547    13.10125        102        178
```

Figure 2.9 shows that the relationship between SBP and weight is very close to linear. The small upward bend at the far left of the graph is mostly due to the outlying observation at the lowest weight and is a warning as to the instability of LOWESS (or any scatterplot smoother) at the edges of the data.

Choice of bandwidth is somewhat subjective. Small bandwidths like 0.05 often give very bumpy curves, which are hard to interpret. At the other extreme, bandwidths too close to one force the curve to be practically a straight line, obviating the advantage of using a scatterplot smoother. See Problem 2.6.

2.4.2.2 Categorical Predictor

With a continuous outcome and a categorical predictor, the usual strategy is to apply the same numerical or graphical methods used for one-variable descriptions of a continuous outcome, but to do so separately within each category of the predictor. As an example, we describe the distribution of SBP in WCGS, within levels of behavior pattern. Table 2.6 shows the most direct way of doing this in Stata. Alternatively, the `table` command can be used to make a more compact display, with command options controlling which statistics are listed. The results are shown in Table 2.7.

Side-by-side boxplots, as shown in Fig. 2.10, are an excellent graphical tool for examining the distribution of SBP in each of the behavior pattern categories and

Table 2.7 Descriptive statistics for systolic blood pressure by behavior pattern

```
. table behpat, contents(mean sbp sd sbp min sbp max sbp)

------------------------------------------------------------
Behaviora |
l Pattern |  mean(sbp)     sd(sbp)    min(sbp)    max(sbp)
----------+-------------------------------------------------
       A1 |   129.2462    15.29221         100         200
       A2 |   129.8891    15.77085         100         212
       B3 |   127.5551    14.78795          98         230
       B4 |   127.1547    13.10125         102         178
------------------------------------------------------------
```

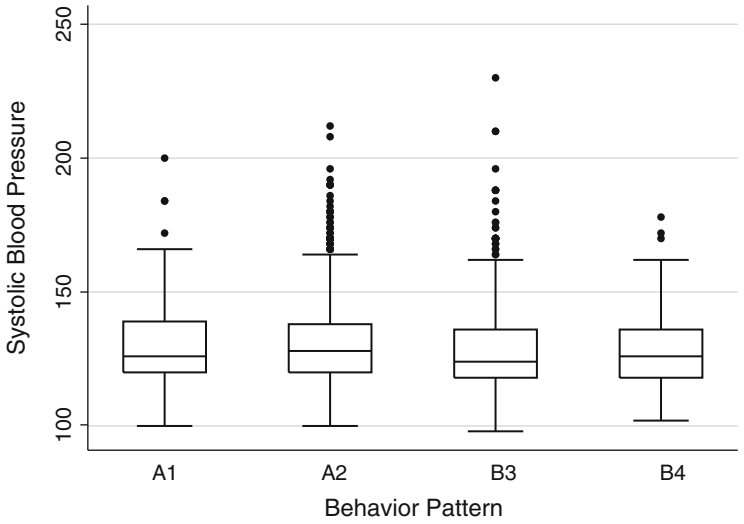

Fig. 2.10 Boxplots of systolic blood pressure by behavior pattern

making comparisons among them. The four boxplots are quite similar. They each have about the same median, interquartile range, and a slight right-skewness. At least on the basis of this figure, there appears to be little relationship between SBP and behavior pattern.

2.4.3 Categorical Outcome Variable

With a categorical outcome variable, the typical method is to tabulate the outcome within levels of the predictor variable. To do so first requires breaking any continuous predictors into categories. Suppose, for example, we wished to treat behavior pattern as the outcome variable and weight as the predictor. We might first divide weight into four categories: $\leq$140 pounds, >140–170, >170–200, and >200. As with histograms, we need enough categories to avoid loss of information, without

Table 2.8 Behavior pattern by weight category

```
. tabulate behpat wghtcat, column

behavioral |
pattern (4 |                    wghtcat
    level) |     < 140    140-170    170-200      > 200 |     Total
-----------+----------------------------------------------+----------
        A1 |        20        125         98         21 |       264
           |      8.62       8.13       8.37       9.86 |      8.37
-----------+----------------------------------------------+----------
        A2 |       100        612        514         99 |      1325
           |     43.10      39.79      43.89      46.48 |     42.01
-----------+----------------------------------------------+----------
        B3 |        90        610        443         73 |      1216
           |     38.79      39.66      37.83      34.27 |     38.55
-----------+----------------------------------------------+----------
        B4 |        22        191        116         20 |       349
           |      9.48      12.42       9.91       9.39 |     11.07
-----------+----------------------------------------------+----------
     Total |       232       1538       1171        213 |      3154
           |    100.00     100.00     100.00     100.00 |    100.00
```

defining categories that include too few observations. Familiar clinical categories are often useful (e.g., glucose <110, 110–125, >125). In Table 2.8, we have requested percentages for each column to facilitate the comparison of the percentages in each behavior pattern between the weight categories. Row percentages or percentages out of the total of 3,154 could also have been requested.

In choosing cutoff points for categorical variables, it is entirely fair to look at the distribution of that variable to try to obtain, for example, roughly equal sample sizes in each of the categories. Splitting the data into $3, 4, 5$, or 10 groups of equal size is a common approach. However, fishing for cutpoints that prove a point is an easy way to arrive at misleading conclusions.

A different strategy with a categorical outcome and a continuous predictor is to "turn the problem around" and treat the continuous variable as the outcome, using the methods of the previous section. If the only goal is to determine whether the two variables are associated, this may suffice. But when the categorical variable is clearly the outcome, this may lead to awkward models and hard-to-interpret conclusions.

2.5 Multivariable Descriptions

Description of more than two or three variables simultaneously quickly becomes difficult. One approach is to look at pairwise associations, e.g., for categorical variables, looking at a series of two-way tables, taking each pair of variables in turn. If a number of the variables are continuous, a correlation matrix (giving all the pairwise correlations) or a scatterplot matrix (giving all the pairwise plots) can be generated. Table 2.9 and Fig. 2.11 show these for the variables SBP, age, weight, and height.The correlation matrix shows that SBP is very weakly correlated with age and weight and essentially uncorrelated with height.

Table 2.9 Correlation matrix for systolic blood pressure, age, weight, and height

```
. correlate sbp age weight height (obs=3154)

             |      sbp      age   weight   height
-------------+------------------------------------
         sbp |   1.0000
         age |   0.1657   1.0000
      weight |   0.2532  -0.0344   1.0000
      height |   0.0184  -0.0954   0.5329   1.0000
```

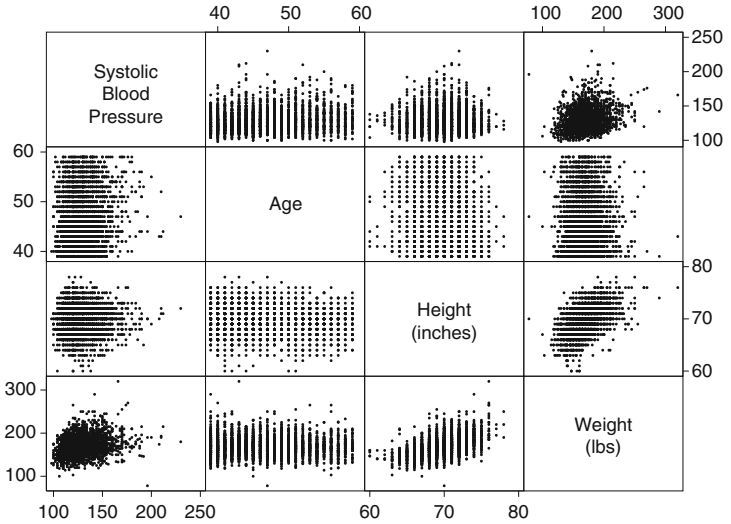

Fig. 2.11 Scatterplot matrix of systolic blood pressure, age, weight, and height

The scatterplot matrix supports the correlation calculation. If one of the variables is clearly the outcome variable, it is useful to list this variable first in the command. That way the first row of the matrix shows the outcome variable on the y-axis, plotted against each of the predictor variables on the x-axis. The matrix of scatterplots for these four variables additionally displays the modest positive correlation between weight and height, indicating the people come in all sizes and shapes!

Multi-way tables that go beyond pairwise relationships can be generated with multiple categorical variables. For example, Table 2.10 shows whether or not the subject had a coronary event (`chd69`), by behavior pattern within weight category. Options in the Stata command are used to obtain the row and column totals. With some study, it is possible to extract information from this three-way table, but it is more difficult than with a one- or two-way table. An advantage of a three-way table is the ability to assess *interaction*, the topic of Sect. 4.6. That is, is the relationship between CHD and behavior pattern the same for each weight category?

Table 2.10 CHD events and behavior pattern by weight category

```
. table chd69 behpat wghtcat, row col
```

	wghtcat and behavioral pattern (4 level)									
	< 140					140-170				
CHD event	A1	A2	B3	B4	Total	A1	A2	B3	B4	Total
no	18	93	84	22	217	115	559	582	184	1,440
yes	2	7	6		15	10	53	28	7	98
Total	20	100	90	22	232	125	612	610	191	1,538

	wghtcat and behavioral pattern (4 level)									
	170-200					> 200				
CHD event	A1	A2	B3	B4	Total	A1	A2	B3	B4	Total
no	81	438	422	108	1,049	20	87	67	17	191
yes	17	76	21	8	122	1	12	6	3	22
Total	98	514	443	116	1,171	21	99	73	20	213

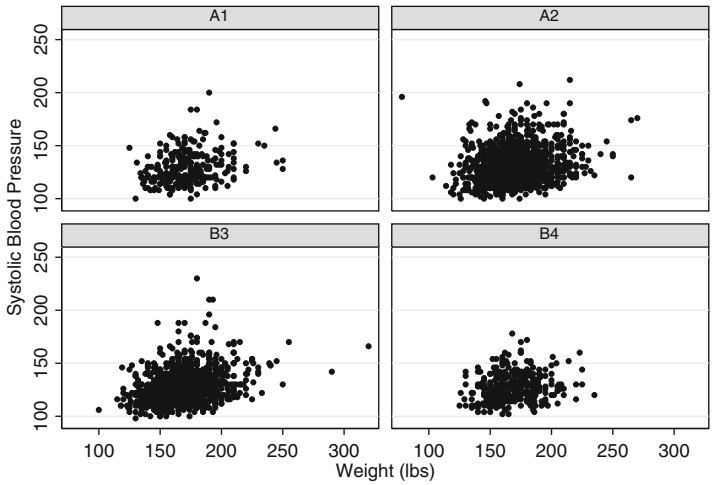

Fig. 2.12 Scatterplot of SBP versus weight by behavior pattern

Analogous graphical displays are also possible. For example, we could look at the relationship between SBP and weight separately by behavior pattern, as displayed in Fig. 2.12. This indicates that the relationship seems to be the same for each behavior pattern, indicating a lack of interaction.

2.6 Summary

Exploratory summaries and graphs are a crucial first step in any data analysis. They provide an opportunity to uncover unusual or anomalous data points which may affect the analysis. Summaries and graphs uncover properties of the data (for instance, skewness) which are useful for informing which model families may fit the data best. Finally, exploring the strength of relationships between variables through graphs provides compelling summaries of the relationships as well as guidance for building regression models.

2.7 Problems

Problem 2.1. Classify each of the following variables as numerical or categorical. Then further classify the numerical variables as continuous or discrete, and the categorical variables as ordinal or nominal.

(1) Gender
(2) Race
(3) Age (in years)
(4) Age in categories (0–20, 21–35, 36–45, 45–60, 60–85, 85+)
(5) Zipcode
(6) Toxicity (mild, moderate, life-threatening, dead)
(7) Number of hospitalizations in the past year
(8) Change in HIV-RNA
(9) Weeks on treatment
(10) Treatment (placebo versus estrogen)

Problem 2.2. Generate pseudo-random data from a normal distribution using a computer program or statistics package. In Stata, this can be done using the `generate` command and the function `invnorm(uniform())`. Now generate a normal Q–Q plot for these data. Do this for several samples of size 10, 50, and 200. How well do the Q–Q plots approximate straight lines? This is valuable practice for judging how well an actual dataset can be expected to approximate a straight line.

Problem 2.3. Generate pseudo-random samples of size 50 from a normal distribution (see Problem 2.2 for how to do this in Stata). Construct histograms of the data using 5, 7, and 15 bins. What do you notice? Do the shapes look like a normal distribution?

Problem 2.4. Warfarin is a drug used to prevent blood clots, for example in patients with irregular heartbeat and after heart surgery. However, too much warfarin can cause unusual bleeding or bruising, so calibration of the dose is important. A study contrasting calibration times (in hours) in two ethnic groups had the following results. For the sample of 19 Caucasians, the times were 2, 4, 6, 7, 8, 9, 10, 10,

12, 14, 16, 19, 21, 24, 26, 30, 35, 44, and 70; for the 18 Asian–Americans, the times were 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 7, 7, 8, 9, 10, 12, 19, and 32.

(1) Display the data numerically to compare the two ethnic groups.
(2) Display the data graphically to compare the two ethnic groups.
(3) Describe the distribution of the data within ethnic group.
(4) Log transform the data and repeat the graphical display. How do the displays with and without log transformation compare?
(5) Can you think of other variables you might want to adjust to help understand the ethnic differences better?

Problem 2.5. The timing of various stages in the contraction of the heart, determined by electro-cardiogram (EKG), can be used to diagnose heart problems. A commonly measured time interval in the contraction of the ventricles is the so-called QRS wave. A study was conducted to see if longer QRS times were related to the ability to induce rapid heart rhythms (called inducible ventricular tachycardia or IVT), which have been associated with adverse outcomes. In a study of 53 subjects, the 18 with IVT had QRS times (in milliseconds) of 70, 75, 86, 90, 96, 102, 110, 114, 116, 117, 120, 130, 136, 142, 145, 152, 170, and 182. The 35 patients without IVT had QRS times of 40, 50, 65, 70, 76, 78, 80, 82, 85, 88, 88, 89, 90, 94, 95, 96, 98, 98, 100, 102, 105, 107, 109, 110, 114, 115, 120, 125, 130, 135, 138, 150, 165, 170, and 180.

(1) Display the data numerically to help understand whether QRS time is related to IVT.
(2) Display the data graphically to help understand whether QRS time is related to IVT.
(3) QRS time is commonly considered as abnormal if the value is greater than 120 ms. Generate a numerical display to help understand if abnormal QRS is related to IVT.
(4) What are the advantages and disadvantages of treating QRS as binary (above 120 ms) instead of continuous?

Problem 2.6. Using the WCGS dataset, generate a LOWESS (or equivalent) scatterplot smooth of SBP versus weight, comparable to Fig. 2.9. Next try the plot with bandwidths of 0.05, 0.15, and 0.50. How do they compare? Which is most useful for judging the linearity or lack of linearity of the relationship? The WCGS data are available at http://www.biostat.ucsf.edu/vgsm.