2

Examining a Single Variable

This chapter introduces the most important statistical graphics to investigate the structure of a single variable. The plots in this chapter can be best categorized by the scale of the variable they display — either continuous or categorical. The latter includes both alphanumerical (nominal) and numerical (possibly ordinal) categorical data.

The most important interactive controls and variations are explained for each plot. Furthermore, the definition and use of highlighting are described.

2.1 Categorical Data

Barcharts and Spineplots

One of the most basic plots is the barchart. As the name implies, a bar is drawn for each category of the variable. The length of each bar is proportional to the number of cases falling into that particular category. Barcharts can be either drawn vertically (what most applications do) or horizontally. Although the vertical layout is probably the more natural way to plot bars, the horizontal layout has the advantage of allowing bar labels to be printed in full length. This is especially useful when working with many categories. Figure 2.1 shows a barchart in a horizontal layout for the variable *Weekday* from the case study in Appendix i, the *Tipping* data. Note that the order of the bars matters a lot in this example. The default order — usually a lexicographic order — would place *Thursday* last, which would make a correct interpretation of the plot unnecessarily complicated.

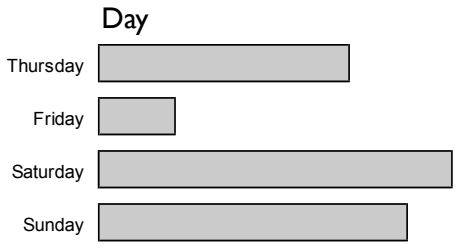

FIGURE 2.1
A barchart for the four days of the week from the Tipping data case study.

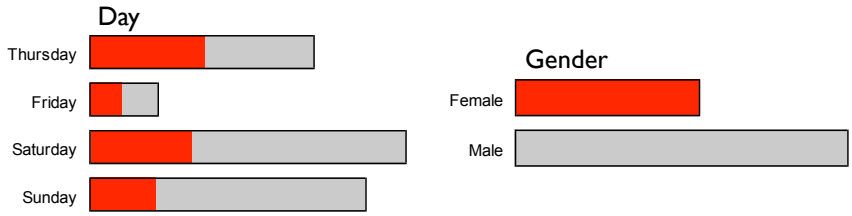

FIGURE 2.2
Highlighting in a barchart.

Adding highlighting to barcharts is straightforward. The barchart of the selected data points is simply drawn on top of the base barchart of the whole sample. Figure 2.2 gives an example of a highlighting in a barchart. All female customers have been selected in the same barchart as shown in Figure 2.1. We see immediately that the distribution of the highlighted cases is not of a different structure than the whole sample, but still, the proportion of females appears to be larger on Thursdays than on Sundays. What makes the comparison of the proportions so difficult? The highlighted part of a bar must be normalized in order to be comparable — a visually challenging task.

Spineplots use normalized bar lengths while the bar widths are proportional to the number of cases in the category. Figure 2.3 shows the data from Figure 2.2, but now the barchart has been switched to a spineplot. The area of the bars is still proportional to the category frequencies. The highlighting proportion can now be compared across all categories since the highlighting direction remains unchanged relative to the barchart and the area of the highlighting is also proportional to the highlighting frequencies. Now we see directly that the proportion of female customers declines monotonously from more than 50% to less than 25% from Thursdays to Sundays. In Mondrian, barcharts and spineplots use the same framework and switching between the two representations takes pressing only a single key.

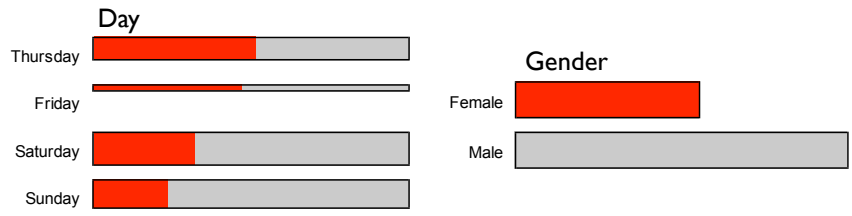

FIGURE 2.3
The same data as in Figure 2.2 in a spineplot.

Looking at absolute and relative amounts of highlighting shows the need for sorting and reordering of barcharts, which is discussed in Chapter 7. The generalization of barcharts and spineplots for more than just one variable, i.e., the mosaic plot, is introduced in Chapter 4.

2.2 Continuous Data

There are far more plots for continuous data than for categorical data. Obviously the amount of information a continuous variable can hold is far greater than a categorical does. Depending on what aspect of a continuous variable is of interest, the one or the other graphic might be the better choice.

Dotplots

Dotplots are a very simple way to plot one-dimensional data. Nevertheless, there are at least three distinct versions. The **standard dotplot** is a scatterplot in one dimension, i.e., a continuous variable is plotted along one axis only. Figure 2.4 shows an example of a standard dotplot for the variable *Tip in USD*. Although this dataset has only 244 observations, we note a strong overplotting for smaller values. No structure is visible for tips less than \$4, whereas the outliers beyond \$6 can be easily spotted.

Jittering is often applied to avoid overplotting in glyph based plots. Jittering is a technique where a small amount of noise — usually white noise, i.e., uniformly distributed random numbers — is added to the data to avoid overplotting. Figure 2.5 shows the same dotplot as in Figure 2.4 now with a small amount of noise added orthogonally to the x axis. In the **jittered dotplot** far more structure is visible in the data even for amounts of less than \$4. The jittering reveals accumulation points at amounts of \$2.00, \$2.50, \$3.00 and \$4.00. Obviously, many customers tend to give a tip of multiples of half a dollar.

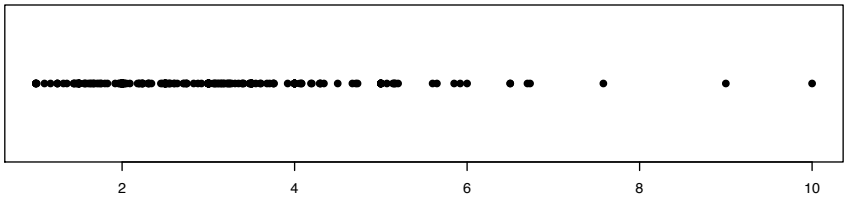

FIGURE 2.4
A standard dotplot for the variable *Tip in USD* of the *Tipping* data.

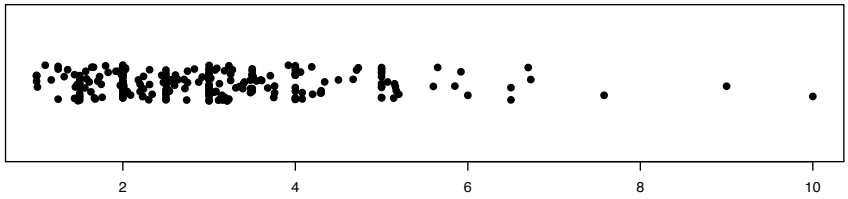

FIGURE 2.5
The same data as in Figure 2.4 with jittering applied.

Although jittering reduces the amount of overplotting considerably, it also introduces a negative effect. The pseudo structure along the virtual y axis may add visual artifacts, as the reader of the graphics will be inclined to interpret the location of points along the y axis as well. A first step to reduce this negative effect is to use so-called **textured dotplots** as introduced by Tukey and Tukey (1990). Textured dotplots use a systematic way to place points side by side to avoid overplotting. There are two drawbacks of such an approach: first, it is impossible to seamlessly switch between the systematic placement and a random placement of points, which is needed for larger datasets. Second, a systematic placement of points needs to have some idea of the density of the variable displayed, which is already a far more complex concept than the initially simple concept of a dotplot. Figure 2.6 shows a dotplot where the amount of jittering is proportional to the data density. This reduces potential visual artifacts due to pseudo structure along the y axis. This variation of a dotplot is probably a data representation giving best insight into the structure of the variable, but also far more complex than a standard dotplot.

A nice overview of the generation of many variations of dotplots can be found in Wilkinson (1999). In an interactive context, the amount of jittering should be controlled interactively and the added noise should be resampled when requested by the user.

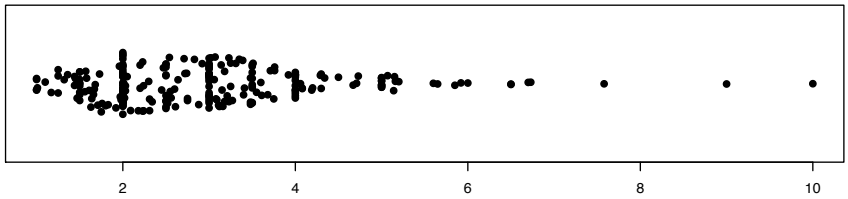

FIGURE 2.6
A dotplot with jitter added which is proportional to the data density.

Boxplots

The boxplot is a graphic which depicts both summary statistics as well as raw data. At the core of a boxplot is the so-called *five number summary*. The five number summary of a variable consists of the minimum, lower hinge, median, upper hinge and maximum. The definitions of the median and the extreme values are well known. The hinges are the medians of the subsamples which are created when dividing the original sample into two parts at the median. Thus, quartiles (0.25 and 0.75 quantiles) and hinges may differ by one index in the sorted sample, which usually does not change the resulting boxplot. Therefore the hinges are often substituted by quartiles for simplicity. Figure 2.7 illustrates all components of a boxplot. The core — the box — is built up by the upper and lower hinges and the median. The difference between the hinges — the so-called *h-spread* — is used to define the *inner fence* and the *outer fence*.* The *whiskers* are drawn from the upper (resp. the lower) hinge to the first value which is no further away from the hinge than 1.5 times the h-spread — the inner fence. All points between inner and outer fence are called *outliers*; all points further away then 3 times the h-spread (the

*The inner and outer fences are never drawn in a boxplot, as they are only imaginary thresholds.

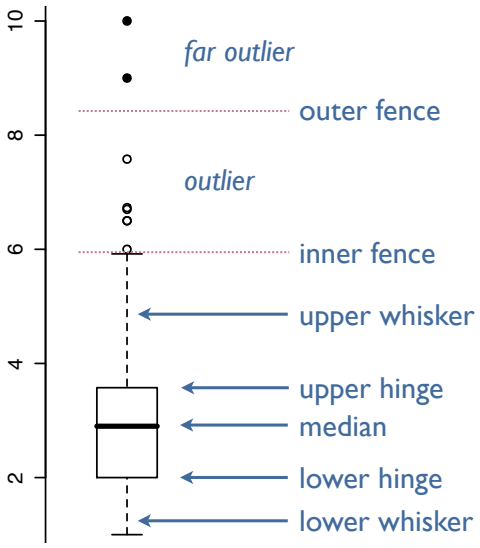

FIGURE 2.7
A boxplot with all its elements annotated.

outer fence) are called *far outliers*. Far outliers are usually marked with a more distinct symbol.

Comparing the boxplot in Figure 2.7 with the three dotplots above shows advantages and disadvantages of this display. The boxplot shows robust measures of location and spread, which gives basic properties of the sample's distribution. These properties are impossible to determine from a dotplot. On the other hand, all data points which are not outliers are represented in an abstract way. Thus it is impossible to see gaps or accumulations in a boxplot, both of which are easy to spot in a dotplot.

Highlighting in a boxplot must respect its special structure made up by summaries and single values. Whereas it is obvious that the glyph of a selected outlier can be highlighted easily, it is not sensible to highlight the box of a boxplot the same way the box in a barchart is highlighted. An example of how highlighting in boxplots can be implemented is shown in Figure 2.8. The upper boxplot shows a base boxplot without any highlighting. In order to be able to plot a highlighted boxplot atop the base plot, the whiskers have been modified and are now light gray boxes extending the inner box of the boxplot. The lower boxplot shows the highlighting. A regular boxplot for the highlighted cases is plotted in the highlighting color atop the base plot. The box of the highlighted boxplot is narrower and slightly transparent such that the parameters of the base boxplot are not obscured.

The definition of a boxplot has some desirable properties, in particular when we assume the data to follow a normal distribution. 50% of the data around the data center lie in the box — regardless of the distribution. For a (standard) normal distribution the quartiles can be found at -0.675 and 0.675 so that the h-spread is approximately 1.35. Adding 1.5 times to the box yields an interval of $[-2.698; 2.698]$. The probability that we observe values outside this interval is $P(x \notin [-2.698; 2.698]) = 0.7\%$. Thus the

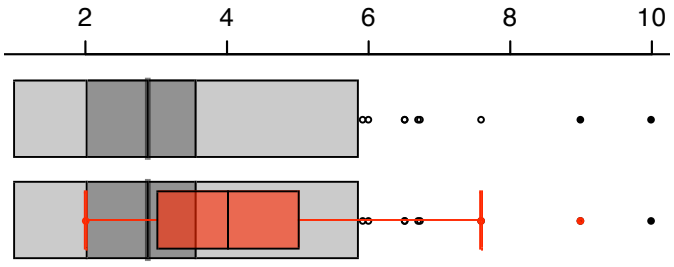

FIGURE 2.8
Highlighting in a boxplot needs a modification of the rendering of the unhighlighted boxplot.

probability that a value is an outlier is just below 1%.[†]

Histograms and Spinograms

Histograms are based on a summary of the sample. For each interval in a set of consecutive intervals, the number of observations falling into that interval is counted. The resulting counts are visualized with bars plotted over the intervals. The area of each bar is proportional to the corresponding count for this interval. The intervals — which are called "bins" — are usually set to have equal width and to be left closed and right open. Figure 2.9 gives an example of two histograms. Both histograms show

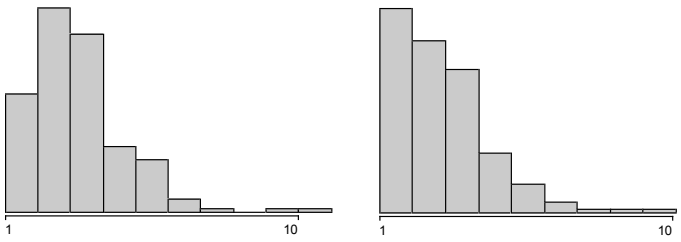

FIGURE 2.9
Two histograms of the variable *Tip in USD*. The left histogram uses bin width of exactly $1, whereas the right histogram uses slightly wider bins of width $1.01. Both start at $1.

the same data. The left histogram uses interval bin width of exactly $1, the right histogram a slightly larger bin width of $1.01. Both histograms start at $1. The apparent shape of the distributions looks quite different. As we see from this example, the parameters that determine a histogram are *bin with* and *anchor point*. For the data in Figure 2.9 a bin width of $1 seems more justifiable, and thus the resulting histogram is probably the better choice.

The dotplots showed accumulations at full and half dollar amounts, which are not visible in either of the histograms of Figure 2.9. To find these accumulations, the bin width has to be set to even smaller amounts than $1.00. Figure 2.10 shows the same data for bin widths $1.00, $0.50 and $0.25. The smaller the bin width, the more apparent are the accumulations at full dollar amounts of $2.00 and $3.00. For a better comparison, the scales of all three histograms have been set to be equal. Since the

[†]An anecdote says that John W. Tukey answered the question why it is 1.5 times the h-spread with: "Because 1 is too small and 2 is too large."

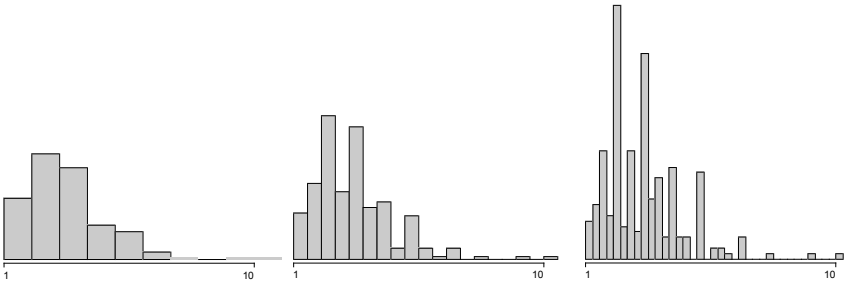

FIGURE 2.10
Three histograms of the variable *Tip in USD*. The chosen bin widths (from left to right) are \$1.00, \$0.50 and \$0.25. All histograms share the same scale and start at \$1.

area of bars is proportional to counts (or proportional to the relative frequency), the sum over the area of all rectangles is the same for all three histograms.

It is obvious from examples in Figure 2.9 and Figure 2.10 how important it is to be able to change bin width and anchor point of a histogram quickly and flexibly. Interactive controls of a histogram must allow us to change the two parameters by a simple mouse drag or keyboard shortcuts. If changing the parameters involves retyping a command and/or creating a whole new plot, the analyst might be inclined to avoid looking at many different views.

Highlighting in histograms can be implemented easily. A highlighted histogram of the selected cases is drawn atop the histogram of all cases. In Figure 2.11 a histogram of the rental price per area is plotted for the data from case study E. All apartments classified to be located in a "good" neighborhood are selected. The immediate question which arises is whether the distribution of the highlighted cases is any different from the distribution of all cases. This question is difficult to answer from the highlighted histogram in Figure 2.11 since we would need to compare the proportions of the highlighted cases across all bars of the histogram, which is visually unfeasible.

One way out is to use the same "trick" as switching from barcharts to spineplots, i.e., all bars are normalized to have the same height, but proportional width. The resulting plot is called a **spinogram** (cf. Hofmann and Theus, under revision). Spineplots have the nice property that highlighted proportions can be compared directly. However, it must be noted that the x axis in a spinogram is no longer linear. It is only piecewise linear within the bars. Although this might be confusing at first sight, it yields two interesting characteristics. Areas where only very few cases have been observed are squeezed together and thus get less visual weight.

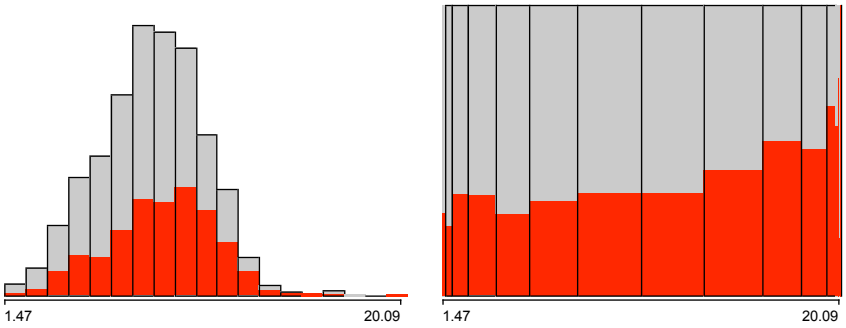

FIGURE 2.11
The variable *Rent per* m^2 from case study E. Left: a histogram with all apartments in a "good" neighborhood highlighted, right: the same data in a spinogram.

Let $\tilde{F}$ denote the empirical distribution function of a variable X, then the x axis in a spinogram is linear in $\tilde{F}^{-1}$. Applications and extensions of spinograms will be further discussed in Section 3.2.

As both views, histograms and spinograms, offer specific insights, it is desirable to switch quickly between the one and the other view, without being forced to create a new plot.

Density Estimation

Histograms are often used to visualize the density of a one-dimensional continuous sample. Figure 2.9 and Figure 2.10 illustrated the strong variation of histograms with the change of the bin width and the anchor point. Histograms are powerful in cases where meaningful class breaks can be defined and classes are used to select intervals and groups in the data. However, they often perform poorly when it comes to the visualization of a distribution.

This drawback was identified long ago, and several strategies have been taken to overcome this problem. One solution is to use so-called **average shifted histograms** or ASHs for short (see Scott, 1992). The idea behind average shifted histograms is quite simple. For a given bin width, the anchor point of a histogram can be shifted within the range of one bin width. Using k different starting points will result in k different histograms. For any given x the average over the k bar heights can then be computed to construct a smoother estimate of the underlying density.

Another method to visualize the density of a variable is to use **kernel density estimators**. The idea behind kernel density estimators is as follows. Given a sample of the size n, each observation contributes $1/n$-th of the density. This contribution to the density is distributed around the

actual observation x_i using a scaled kernel function $k(x)$ at point x_i. For a given x the resulting density estimate can then be summed up over all contributions $k_{x_i}(x)$ each centered around the n x_i's, yielding

$$\hat{f}(x) = \frac{1}{cn} \sum_{i=1}^{n} k\left(\frac{x - x_i}{c}\right) \quad \text{for} \quad k(x) = k(-x). \tag{2.1}$$

For all kernel functions k

$$\int_{-\infty}^{\infty} k(x)dx = 1, \quad \int_{-\infty}^{\infty} k^2(x)dx < \infty, \quad \left|\frac{k(x)}{x}\right| \to 0 \quad \text{for } |x| \to \infty\,. \tag{2.2}$$

Figure 2.12 illustrates how a kernel density estimate is assembled from n kernel functions for the x_i using a normal density as kernel. The $n = 244$ cases of the variable *Tip in USD* with their corresponding kernels $k(x_i)$ are plotted in blue. Summing these functions up for each x gives the resulting density estimate plotted in purple. Note that the functions of the blue and the purple curves are drawn on a different scale such that the kernel functions are visible. Various kernels can be used such as rectangular, triangular or normal.

It can be shown that ASHs converge for $k \to \infty$ toward a kernel density estimate with a triangular kernel (see Venables and Ripley, 1999).

Figure 2.13 shows the three histograms from Figure 2.10. Each histogram has a kernel density estimate superposed which uses a bandwidth c equal to the bin width of the underlying histogram. The leftmost estimate is clearly oversmoothed and cannot capture the structure of the variable, whereas the rightmost estimate looks quite rough and is thus

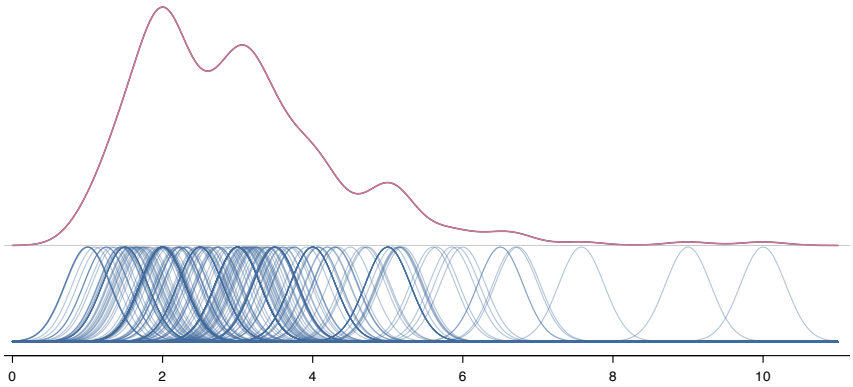

FIGURE 2.12
Illustration of how a kernel density is assembled out of the n contributing points x_i.

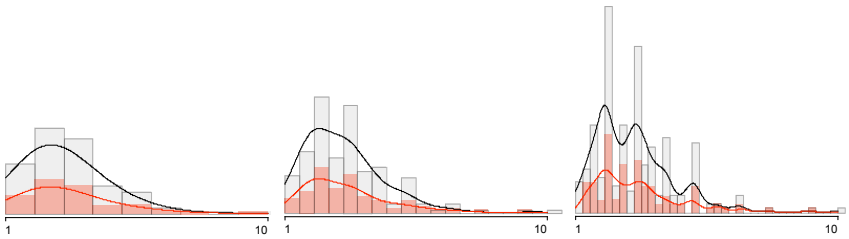

FIGURE 2.13
The same histograms as in Figure 2.10 now with kernel density estimates superposed. Male smoker parties are selected.

not satisfying. This trade-off — also generally referred to as bias-variance trade-off — can be best investigated when the bandwidth c of the density estimator can be varied interactively such that the analyst can see the change of the estimate instantaneously. For the data in Figure 2.13 there seems to be no 'best' bandwidth which captures the composite density.

CD-Plot

Although the spinogram is an efficient, area proportional display to visualize the conditional distribution of a subgroup of a continuous variable, the transformed x-axis of the spinogram can be difficult to interpret in some cases. The CD-plot visualizes the conditional distribution (CD) by

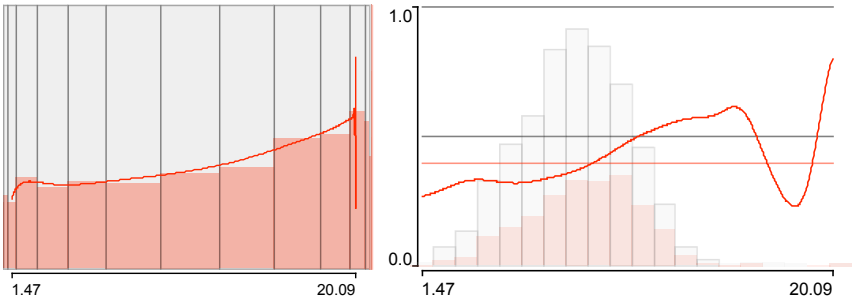

FIGURE 2.14
The variable *Rent per* m^2 from case study E. Left: a spinogram with all apartments in a "good" neighborhood highlighted and a density estimate superimposed, right: the same data in a CD-plot. The CD-plot preserves the scale, whereas the spinogram focuses on intervals with a significant signal.

setting the density estimate of the selected subgroup in relation to the density estimate of the complete sample. The histogram is used just as a backdrop for orientation.

Figure 2.14 shows a CD-plot for the same data as in Figure 2.11. The trend in the data is the same as for the spinogram. The strong variation of the estimate for prices between €15 and €20 is due to the small number of points on which the estimate is based in this interval. This statistically insignificant information is avoided in spinograms since areas of very low density are squeezed to intervals of almost zero size.

2.3 Transforming Data

Continuous Data

Transforming data can have many motivations. A method requires normally distributed data to perform correctly, an extreme skewness of a distribution squeezes 99% of a variable's data onto 1% of the range of the variable, or the data simply needs a transformation into an established, more readily interpretable scale.

Transforming a variable is one of the earliest procedures found in dynamic graphics. The Box-Cox transformation defined by

$$x_{BC}(\lambda, \alpha) = \begin{cases} \dfrac{(x+\alpha)^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(x+\alpha) & \text{for } \lambda = 0 \end{cases} \tag{2.3}$$

is the most common transformation for continuous variables and generalizes a simple logarithmic transformation to a more general power transformation.

A **qqplot** is often used to verify to which degree a sample follows a specific distribution. They plot the empirical quantile $x_{(i)}$ of the ordered sample against its theoretical quantile, e.g., for a standard normal distribution $z(\frac{i}{n+1})$. If a sample follows the theoretical distribution, all points in the scatterplot fall approximately on a line.

Figure 2.15 shows a qqplot for the variable *Tiprate* from case study i. Obviously the data are not normal, as several points deviate strongly from the line.

There are several approaches to make the data more normal. The simplest — and often used — is to take logs. In many situations the technical background even allows us to interpret the logarithm of a quantity (e.g., the quantification of sound waves).

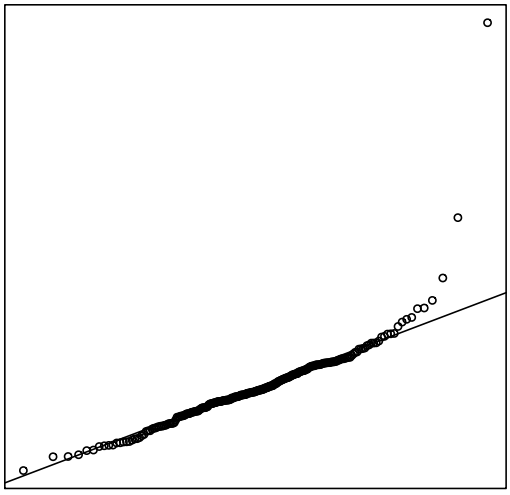

FIGURE 2.15
qqplot of the tiprate from case study i. (axes are omitted as they cannot aid interpretation).

The resulting qqplot is shown in Figure 2.16 left. As the result is not satisfying, a Box-Cox transformation with an optimized λ-value might be more appropriate. Using a dynamic transformation in DataDesk it is easy to find that for $\lambda = 0.155$, the variable is no longer skewed and the shape is very close to a normal distribution (λ has actually been chosen to match the skewness $\gamma_1 = 0$ and a kurtosis $\beta_2 = 3$ of a standard normal distribution). The resulting qqplot in Figure 2.16 (center) shows an improvement over the plain log-transformation, but many points still deviate from the line.

Neither the log-tranformation nor the Box-Cox transformation gives satisfying results in the qqplot. There is one more option left: removal of outliers. Figure 2.16 right shows the qqplot with the three largest values removed — the corresponding boxplot of the variable shows 4 outliers with the smallest of them being only slightly above the upper whisker. The qqplot confirms that this is the best solution of all three approaches. There is no general rule as to how to handle skewed or non-normal data. If normality is a prerequisite of a method a Box-Cox transformation or the removal of outliers might do the job. Both solutions have their problems. Any conclusion drawn from an analysis of a transformed variable must be retranslated into the original domain — which is usually not an easy task. A special handling of outliers, be it a complete removal, or just visual suppression such as hot-selection or shadowing, must have a cogent motivation. At any rate, transformations of data are usually part of a data preprocessing step that might precede a data analysis. Also it can be motivated by initial findings in a data analysis which revealed yet undiscovered problems in the dataset.

Default transformations which standardize data either by mean and standard deviation or onto a $[0, 1]$ range should generally be avoided, as they put all data on scales which can no longer be interpreted.

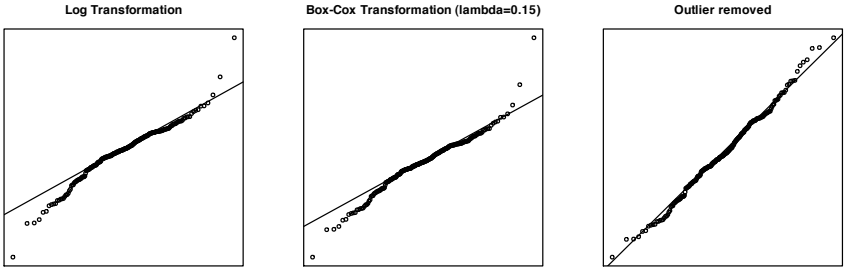

FIGURE 2.16
Three approaches to achieve normality: log-transformation (left), Box-Cox transformation with optimized λ (middle) and a simple exclusion of the three biggest outliers (right).

Categorical Data

At a first sight there might not be much to transform in a categorical variable. But even for categorical data there is a problem similar to skewness. Ordinal, numerical variables such as "number of persons in the family" tend to have the majority of observations distributed among only a few classes.

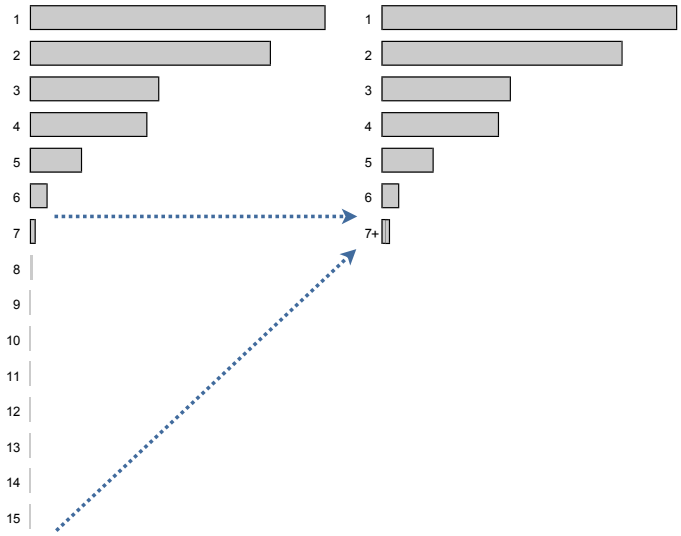

FIGURE 2.17
In the barchart for *Number of Persons in Household*, only 0.5% of all cases make up more than half of the categories.

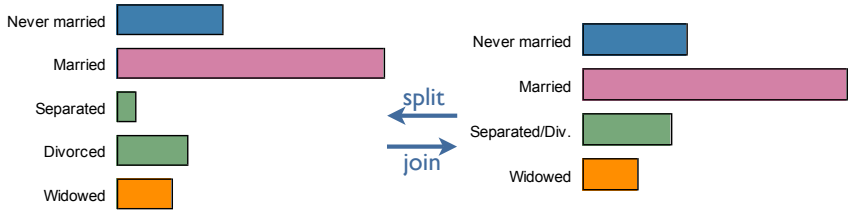

FIGURE 2.18
Joining and splitting classes can be useful even for just a few categories.

Figure 2.17 shows the variable *Number of Persons in Household* for the *Current Population Survey '95* which covers 63,756 of households in the U.S. In this example, only 0.5% of the data make up 8 out of 15 categories. In order to avoid clutter, all categories bigger than 6 can be joined to a new class "7+" as indicated in Figure 2.17 right. One might think that summarizing classes 7 to 15 is a typical preprocessing job. On the other side, it is very efficient to have the ability to decide on such an operation on the fly in an interactive system. Nevertheless, all operations which change data can be a source of errors or at least of misinterpretations and thus should only be applied with much care.

Joining and splitting classes can be very effective even with a few categories. For the same census data used in Figure 2.17 the variable *Marital Status* is shown in a barchart in Figure 2.18. The left barchart shows the two categories *Separated* and *Divorced* as two distinct classes, whereas the right barchart shows the joined version. The two barcharts are not really very different, but all analyses and graphics which are split by, i.e., conditioned on, the 4 or 5 groups of the variable might change substantially.

2.4 Weighted Plots

Basically, one can distinguish three motivations for weighted data. The first is a technical motivation. Whenever we look at purely categorical data, it is not necessary to supply a dataset case by case. A breakdown summary can capture the dataset without loss of any information. Figure 2.19 shows the first 6 lines of the raw data of the *Titanic* dataset, case by case. All six lines are identical as the group of adult male first class passengers who survived has size 140. In this format, the whole dataset has 2,201 entries. The far more efficient version of this information is the summarized data table shown in Figure 2.20. In this representation, the

Class	Age	Gender	Survived
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes
First	Adult	Male	Yes
...			

FIGURE 2.19
The first 6 lines of the *Titanic* dataset in raw format, case by case.

dataset has an extra column specifying the size of each group. Since *Class* has four categories, *Age*, *Gender* and *Survived* two each, the dataset will have at most $4 \times 2 \times 2 = 32$ lines. Because 8 of the 32 combinations of the variables do not occur in the data, the dataset can be reduced to 24 lines.

Summarized data tables can be obtained via database queries. A database containing the *Titanic* data case by case can by queried with the simple SQL command.

```
SELECT Class,
       Age,
       Gender,
       Survived,
       count(*)
 FROM Titanic
 GROUP BY Class,
          Age,
          Gender,
          Survived
```

For the *Titanic* dataset, it makes no real difference whether we handle the 2,201 cases in a raw format or the summarized version as long as the software is capable of handling both formats. However, as datasets

Class	Age	Gender	Survived	Count
First	Adult	Female	No	4
First	Adult	Female	Yes	140
First	Adult	Male	No	118
First	Adult	Male	Yes	57
First	Child	Female	Yes	1
First	Child	Female	No	0
...				

FIGURE 2.20
The first 6 lines of the *Titanic* dataset in summarized form.

get really large this difference can become dramatic. The second situation in which weights are introduced is when sampling unequally from a population. Statistics and graphics must then account for the weights. A third reason to use weights is a change of the sampling population. For example a dataset on cancer rates measured on a county level might be weighted with the population of a county in order to switch from the distribution of rates within counties to the distribution of actual cancer cases.

How can statistical graphics incorporate weights? The modification is quite simple for area-based plots displaying counts. Whereas in an unweighted plot bar sizes are proportional to the count in a class, a weighted plot has bar sizes proportional to the sum of the weights in a class. This modification covers plots such as barcharts, histograms and mosaic plots. Glyph-based plots need different modifications. In a scatterplot the point sizes might be adjusted according to the weights. However, this may lead to overplotting and large differences in individual weights could obscure the scatterplot as a whole.

Figure 2.21 gives an example of a weighted histogram compared with the unweighted histogram. The left histogram shows the percentage of votes for G.W. Bush in the 2004 presidential election for the 65 counties in Florida (cf. case study I). The right histogram shows the same plot now weighted by the number of votes in each county. Note that as we change the population from counties to voters — the e.g., rightmost bar ranging from 75 to 80 percent corresponds to 5 counties, this bar represents roughly a quarter million votes (out of almost 7.5 million votes) in the right histogram — the y-scales are no longer comparable. Switching from unweighted to weighted histograms, i.e., from counties to voters, shows that Bush's support was stronger in the less populated counties since high percentages are downweighted and low percentages are upweighted.

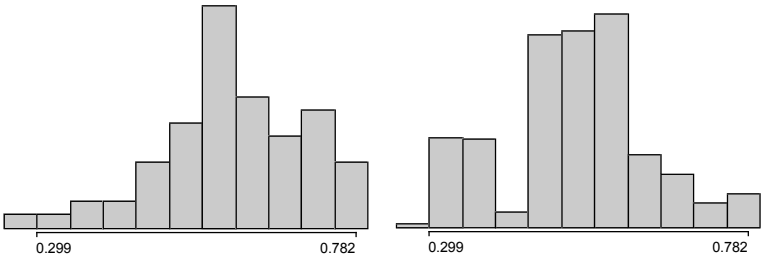

FIGURE 2.21
Histograms of the percentage votes for G.W. Bush in the 2004 presidential election in Florida. Left: unweighted counts by county, Right: percentages weighted by the number of voters.

Exercises

2.1. Barcharts and Spineplots
For the *Tipping* data in case study i

(a) Use barcharts and spineplots to investigate on what days smoking parties are most common.

(b) What problem arises when we look at the size of smoking parties?

2.2. Dotplots, Boxplots and qqplots
For the *Tipping* data in case study i

(a) Compare the benefits of dotplots and qqplots for the variable *Tip in USD*.

(b) What is the percentage of points classified by a boxplot to be an outlier if the underlying distribution is assumed to be standard log-normal?

(c) Can we expect to observe outliers in a boxplot at the steep side of a skewed distribution?

2.3. Histograms & Spinograms
Recreate the graphics from Figure 2.11. What can be said about the rental prices of apartments in buildings built before World War II?

2.4. Density Estimators

(a) Create a histogram with a density estimate for the tiprate from case study i. How do the outliers influence the estimate?

(b) What does the shape of a density estimate look like for the bandwidth $c \to \infty$?

2.5. Transformations
For the *Tipping* data in case study i

(a) Draw a qqplot for the tiprate with the three largest and the smallest values removed. Does the qqplot improve over Figure 2.16?

(b) Investigate the four outliers in the boxplot for *Tiprate*. Why could these cases be treated separately or even neglected?

2.6. Weighted Plots

(a) Create the two graphs in Figure 2.21 of the Florida election data for John F. Kerry. Do these graphs yield a consistent interpretation to what we learned from Figure 2.21?

(b) Create a summarized version of the data from case study B. How many lines has the file?