Chapter 5
Exploratory Data Analysis

5.1 Introduction

Exploratory data analysis is the process by which a person manipulates data with the goal of learning about general patterns or tendencies and finding specific occurrences that deviate from the general patterns. Much like a detective explores a crime scene, collects evidence and draws conclusions, a statistician explores data using graphical displays and suitable summaries to draw conclusions about the main message of the data.

John Tukey and other statisticians have devised a collection of methods helpful in exploring data. Although the specific data analysis techniques are useful, exploratory data analysis is more than the methods – it represents an attitude or philosophy about how data should be explored. Tukey makes a clear distinction between *confirmatory* data analysis, where one is primarily interested in drawing inferential conclusions, and *exploratory* methods, where one is placing few assumptions on the distributional shape of the data and simply looking for interesting patterns. Good references on exploratory methods are Tukey [47] and Hoaglin et al. [22].

There are four general themes of exploratory data analysis, namely *Revelation*, *Resistance*, *Residuals*, and *Reexpression*, collectively called the *four R's*. There is a focus on *revelation*, the use of suitable graphical displays in looking for patterns in data. It is desirable to use *resistant* methods – these methods are relatively insensitive to extreme observations that deviate from the general patterns. When we fit simple models such as a line, often the main message is not the fitted line, but rather the *residuals*, the deviations of the data from the line. By looking at residuals, we often learn about data patterns that are difficult to see by the initial data displays. Last, in many situations, it can be difficult to see patterns due to the particular measuring scale of the data. Often there is a need to *reexpress* or change the scale of the data. Well-chosen reexpressions, such as a log or square root, make it easier

J. Albert and M. Rizzo, *R by Example*, Use R, DOI 10.1007/978-1-4614-1365-3_5,

to see general patterns and find suitable data summaries. In the following example, we illustrate each of the four "R themes" in exploratory work.

5.2 Meet the Data

Example 5.1 (Ratings of colleges).

It can be difficult for an American high school student to choose a college. To help in this college selection process, *U.S. News and World Report* (`http://www.usnews.com`) prepares a yearly guide *America's Best Colleges*. The 2009 guide ranks all of the colleges in the United States with respect to a number of different criteria. The dataset `college.txt` contains data in the guide collected from a group of "National Universities." These are schools in the United States that offer a range of degrees both at the undergraduate and graduate levels. The following variables are collected for each college:

a. School – the name of the college
b. Tier – the rank of the college into one of four tiers
c. Retention – the percentage of freshmen who return to the school the following year
d. Grad.rate – the percentage of freshman who graduate in a period of six years
e. Pct.20 – the percentage of classes with 20 or fewer students
f. Pct.50 – the percentage of classes with 50 or more students
g. Full.time – the percentage of faculty who are hired full-time
h. Top.10 – the percentage of incoming students who were in the top ten percent of their high school class
i. Accept.rate – the acceptance rate of students who apply to the college
j. Alumni.giving – the percentage of alumni from the college who contribute financially

We begin by loading the dataset into R using the function `read.table` and saving it in the data frame `dat`. Note that the `sep` argument indicates there are tabs separating the columns in the data file.

```
> dat = read.table("college.txt", header=TRUE, sep="\t")
```

There are some colleges where not all of the data were collected. The R function `complete.cases` will identify the colleges where all of the variables have been collected and the `subset` function is used to create a new data frame `college` containing only the colleges with "complete" data.

```
> college = subset(dat, complete.cases(dat))
```

5.3 Comparing Distributions

One measure of the quality of a college is the variable `Retention`, the percentage of freshmen who return to the college the following year. We wish to display the distribution of the retention rates and compare the distribution of rates across different college subgroups.

5.3.1 Stripcharts

One basic graph of the retention rates is a stripchart or one-dimensional scatterplot constructed using the `stripchart` function. Using the `method = "stack"` option, the dots in the graph will be stacked, and the option `pch = 19` will use solid dots as the plotting character.

```
> stripchart(college$Retention, method="stack", pch=19,
+     xlab="Retention Percentage")
```

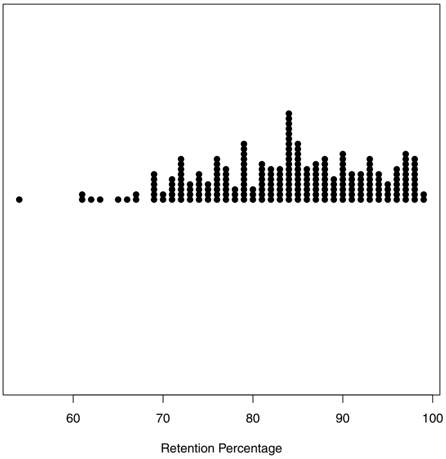


Fig. 5.1 Stripchart of the retention percentages of all National universities.

From Figure 5.1, we see much variability in the retention rates from 55% to near 100% and wonder which variables are helpful in explaining this variation. One of the general measures of a school's quality is its Tier (either 1, 2, 3, or 4) and a next step might be to construct parallel stripcharts of the retention rates by Tier. This graph is constructed by a slight variation of `stripchart` – the argument `Retention` ∼ `Tier` indicates that we wish separate displays of retention for each of the four tiers. Note that we don't use the `college$Retention` syntax, since we have indicated by the `data=college` argument that the data frame `college` is being used.

```
> stripchart(Retention ~ Tier, method="stack", pch=19,
+     xlab="Retention Percentage",
+     ylab="Tier", xlim=c(50, 100), data=college)
```

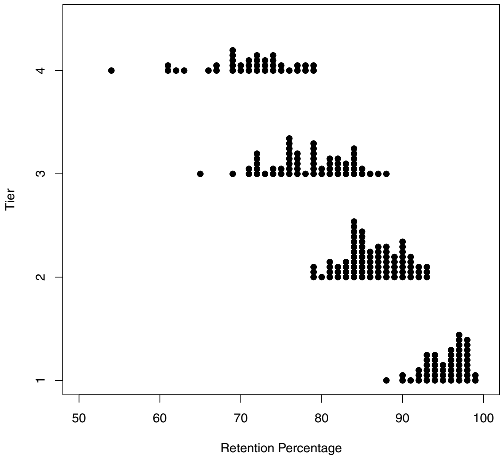


Fig. 5.2 Parallel stripcharts of the retention percentages of all National universities grouped by tier.

It is clear from Figure 5.2 that the retention percentage differs among the four groups of colleges. For the Tier 1 schools, the percentages primarily fall between 90 and 100; in contrast, most of the retention percentages of the Tier 4 schools fall between 65 and 80.

5.3.2 Identifying outliers

The parallel stripcharts are helpful in seeing the general distribution of retention percentages for each tier. Also from the graph, we notice a few schools with retention percentages that seem set apart from the other colleges in the same tier. The `identify` function is helpful in identifying the schools with these unusual percentages. In this function, we give the x and y variables of the plot, indicate by the `n=2` option that we wish to identify two points, and the `labels=college$School` option indicates that we wish to label the points by the school names. When this function is executed, a cross-hair will appear over the graph and one moves the mouse and clicks at the locations of the two outliers. At each click, the name of the school will appear next to the plotting point. In the R console window, the row numbers of the data frame corresponding to these two schools are displayed.

```
> identify(college$Retention, college$Tier, n=2,
+   labels=college$School)
[1] 158 211
```

We see in Figure 5.3 that the two schools with unusually small retention percentages (relative to the other schools in the same tier) are Bridgeport in Tier 4 and South Carolina State in Tier 3.

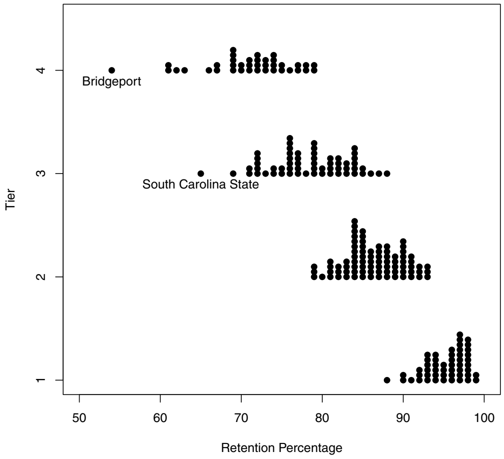


Fig. 5.3 Parallel stripcharts of the retention percentages of all National universities grouped by tier. Two outliers are identified by the school name.

5.3.3 Five-number summaries and boxplots

The parallel stripchart display shows differences in the retention percentages among the four tiers of colleges, and a next step is to summarize these differences. A good collection of summaries of a dataset is the median, the lower and upper quartiles, and the low and high values. This group of summaries is called (for obvious reasons) the *five-number summary* and a *boxplot* is a graph of these five numbers. The `boxplot` function will compute five-number summaries for each group and display parallel boxplots (see Figure 5.4). As in the `stripchart` function, the `Retention` ∼ `Tier` formula indicates that we wish to construct boxplots of retention by tier, and the `horizontal=TRUE` argument indicates the boxplots will be displayed in a horizontal style. As in the stripchart function, the `data=college` argument indicates that the variables are part of the data frame `college`. The resulting boxplot display is shown in Figure 5.4. The locations of the median retention values for the four tiers are shown by the dark lines in the boxes, and the spreads of the four graphs are reflected by the widths of the boxes. Using an EDA rule for flagging outliers, the display shows four schools (indicated by separate points) whose retention percentages are unusually small for their associated tiers.

```
> b.output = boxplot(Retention ~ Tier, data=college, horizontal=TRUE,
+   ylab="Tier", xlab="Retention")
```

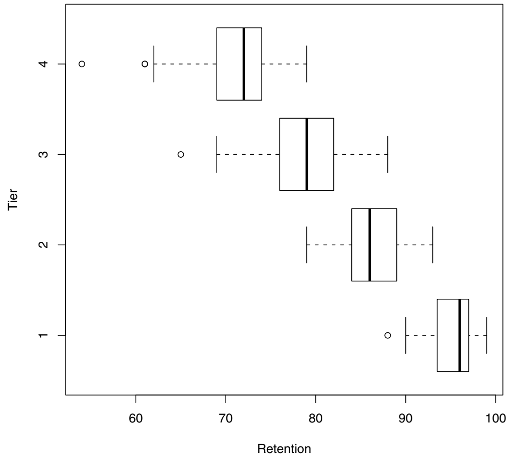


Fig. 5.4 Boxplots of the National university retention percentages grouped by Tier.

The output of `boxplot` has been saved to the variable `b.output`, a list with the different components of the boxplot computations. One can display the five-number summaries by the `stats` component of `b.output`:

```
> b.output$stats
     [,1] [,2] [,3] [,4]
[1,] 90.0   79   69   62
[2,] 93.5   84   76   69
[3,] 96.0   86   79   72
[4,] 97.0   89   82   74
[5,] 99.0   93   88   79
attr(,"class")
        1
"integer"
```

A column of b.output$stats corresponds to the five-number summary of the retention rates for a particular tier. We see the five-number summaries of tiers 3 and 4 are respectively (69, 76, 79, 82, 88) and (62, 69, 72, 74, 79). One can measure the spread of the two groups of data by the quartile spread, the distance between the two quartiles. The quartile spread of the retentions for Tier 3 schools is $82 - 76 = 6$ and the quartile spread of the Tier 4 schools is $74 - 69 = 5$. Since the spreads of the two groups are similar, one can compare the medians – the median for the Tier 3 retentions is 79 and the median for the Tier 4 retentions is 72. We observe that the retention percentages tend to be $79 - 72 = 7$ points higher for Tier 3 than for Tier 4. In addition to the five-number summaries, information about the outliers is also stored. The out and group components of b.output give the outlying values and their corresponding groups.

```
> b.output$out
[1] 88 65 61 61 54

> b.output$group
[1] 1 3 4 4 4
```

From Figure 5.4, there were two visible outliers in tier 4, but the output indicates that there are actually three outliers in this tier, corresponding to retention percentages of 54, 61, and 61.

5.4 Relationships Between Variables

5.4.1 Scatterplot and a resistant line

Since it is reasonable to believe a school's first-year retention percentage will affect its graduation percentage, we next look at the relationship between the variables Retention and Grad.rate. Using the plot function, we construct a scatterplot and the resulting graph is shown in Figure 5.5. As expected, we see a strong positive association between first-year retention rate and the graduation rate.

```
> plot(college$Retention, college$Grad.rate,
+  xlab="Retention", ylab="Graduation Rate")
```

For exploratory work, it is useful to fit a line that is resistant or not sensitive to outlying points. One fitting method of this type is Tukey's "resistant line" implemented by the function line. Essentially, the resistant line procedure divides the scatterplot into left, middle, and right regions, computes resistant summary points for each region, and finds a line from the summary points. We fit this resistant line to these data and the fitting calculations are stored in the variable fit. In particular, the coefficients of the fitted line are stored in fit$coef:

```
> fit = line(college$Retention, college$Grad.rate)
> coef(fit)
[1] -83.631579   1.789474
```

The fitted line is given by

$$\text{Graduation Rate} = -83.63 + 1.79 \times \text{Retention Rate}.$$

The slope of this line is 1.79 – for every one percent increase in the retention rate, the average graduation rate increases by 1.79%. The line on the scatterplot is added by the `abline` function in Figure 5.5.

```
> abline(coef(fit))
```

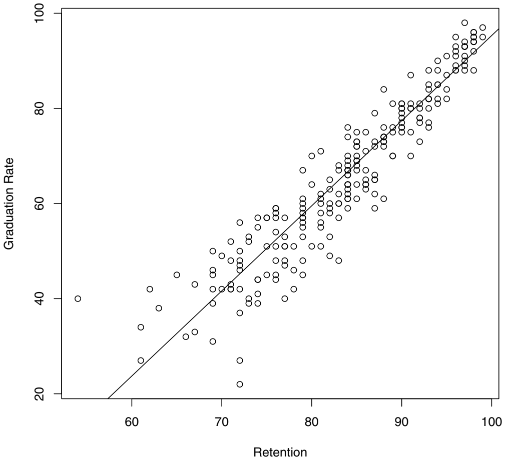


Fig. 5.5 scatterplot of the graduation rates and retention rates for the National universities. A resistant best-fitting line is placed on top of the graph.

5.4.2 Plotting residuals and identifying outliers

In exploratory work, we wish to look beyond the obvious relationship between retention and graduation rates by examining schools deviating from the general straight-line pattern. We look further by considering the *residuals*, the differences between the actual graduation rates and the values predicted from the fitted resistant line. The set of residuals are stored in the list element `fit$residuals` and the `plot` function is used to construct a scatterplot of residuals against the retention rates in Figure 5.6. A horizontal line at zero is added using `abline` to help in interpreting this plot. Positive residual values correspond to observed graduation rates that are larger than the values predicted from the straight-line relation, and negative residuals correspond to graduate rates smaller than the predicted rates.

```
> plot(college$Retention, fit$residuals,
+  xlab="Retention", ylab="Residual")
> abline(h=0)
```

We learn some new things from inspecting this residual plot. There is a fan-shaped appearance in the graph, indicating that the spread of the residuals is higher for low retention schools than for high retention schools. Most of the residuals fall between −20 and 20 percentage points, indicating that the observed and predicted graduation rates fall within 20 points for most schools. We do notice two unusually large residuals, and we identify and label these residuals using the `identify` function.

```
> identify(college$Retention, fit$residuals, n=2,
+  labels=college$School)
```

When this function is executed, a crosshair will appear on the graph. One clicks at the locations of the two large residuals and the names of the schools appear next to the plotting points. (See Figure 5.6.) The two large residuals correspond to the schools Bridgeport and New Orleans. Although Bridgeport has a relatively low retention percentage, it has a large positive residual which indicates that its graduation percentage is large given its retention percentage. Perhaps Bridgeport's actual retention percentage is higher than what was recorded in this dataset. In contrast, New Orleans has a large negative residual. This school's graduation percentage is lower than one would predict from its retention percentage. This suggests that another variable (a so-called *lurking variable*) may explain New Orleans' low graduation percentage.

5.5 Time Series Data

5.5.1 Scatterplot, least-squares line, and residuals

Example 5.2 (Enrollment growth at a university).

Bowling Green State University celebrated its centennial in 2010 and it published online its enrollment counts for the years 1914 through 2008. The dataset "bgsu.txt" contains the enrollment counts for the significant growth years 1955 through 1970. We read in the dataset and use the `plot` function to construct a scatterplot of `Enrollment` against `Year`. (See Figure 5.7.)

```
> bgsu = read.table("bgsu.txt", header=TRUE, sep="\t")
> plot(bgsu$Year, bgsu$Enrollment}
```

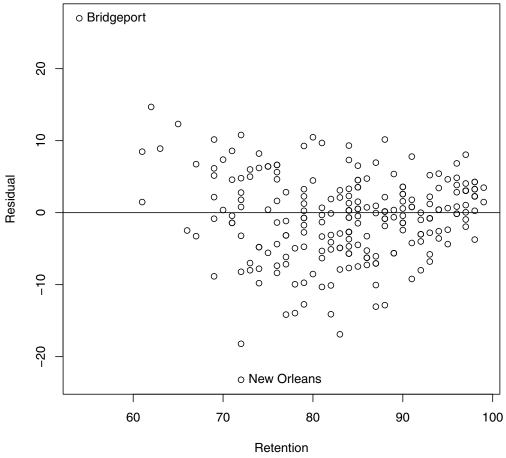


Fig. 5.6 Plot of the residuals of a resistant fit to the graduation percentages by the retention percentages for the National universities. Two unusually large residuals are labeled with the corresponding college name.

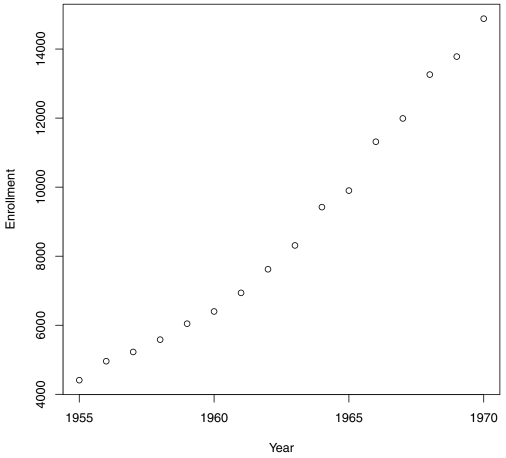
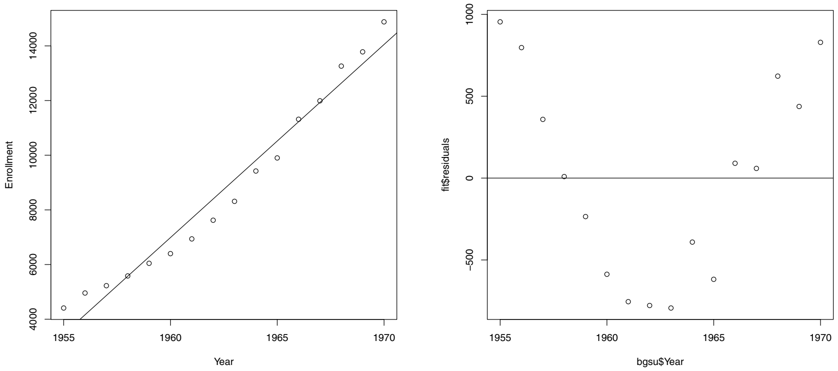


Fig. 5.7 scatterplot of BGSU enrollment against year for the growth period 1955 to 1970.

To help us understand the pattern of growth of enrollment, we fit a line. The `lm` function is used to fit a least-squares line with the calculations stored in the variable `fit`. The fitted line is placed on the scatterplot using the `abline` function with the argument `fit`. The vector of residuals is stored in the component `residuals` of `fit` and the `plot` function is used to construct a scatterplot of the residuals against year. The `abline` function is used with the `h=0` argument to add a horizontal line at zero to the residual plot. Figure 5.8 shows the two plots.

```
> fit = lm(Enrollment ~ Year, data=bgsu)
> abline(fit)
> plot(bgsu$Year, fit$residuals)
> abline(h=0)
```



(a) (b)

Fig. 5.8 Least-squares fit to enrollment data (a) and residual plot (b). There is a clear curvature pattern to the residuals, indicating that the enrollment is not increasing in a linear fashion.

Looking at the residual graph, there is a clear curvature pattern in the residuals, indicating that BGSU's enrollment is not increasing in a linear way. An alternative model may better describe the enrollment growth.

5.5.2 Transforming by a logarithm and fitting a line

Suppose instead that the BGSU enrollment is increasing exponentially. This means that, for some constants a and b, the enrollment follows the relationship

$$Enrollment = a\exp(bYear).$$

If we take the logarithm of both sides of the equation, we obtain the equivalent linear relationship between the log of enrollment and year

$$\log Enrollment = \log a + bYear.$$

We can find suitable constants a and b by fitting a line to the (Year, log Enrollment) data. In the following R code, we define a new variable `log.Enrollment` containing the log enrollment values.

```
> bgsu$log.Enrollment = log(bgsu$Enrollment)
```

℞ **5.1** *The syntax* `bgsu$log.Enrollment` *on the left side of the assignment creates a new variable* `log.Enrollment` *in the* `bgsu` *data frame. The logarithm of enrollment is then assigned to this new variable.*

We construct a scatterplot of the reexpressed data against year, and use the `lm` function to fit a line to these data. Figure 5.9 displays the least-squares fit to the log enrollment data, and plots the corresponding residuals. Generally, it seems that a linear pattern is a closer fit to the log enrollment than for the enrollment. Looking at the residual graph, there is a "high, low, high, low" pattern in the residuals as one looks from left from right, but we do not see the strong curvature pattern that we saw in the residuals from the linear fit to the enrollment.

```
> plot(bgsu$Year, bgsu$log.Enrollment)
> fit2 = lm(log.Enrollment ~ Year, data=bgsu)
> fit2$coef
  (Intercept)          Year
-153.25703366    0.08268126
> abline(fit2)
> plot(bgsu$Year, fit2$residuals)
> abline(h=0)
```

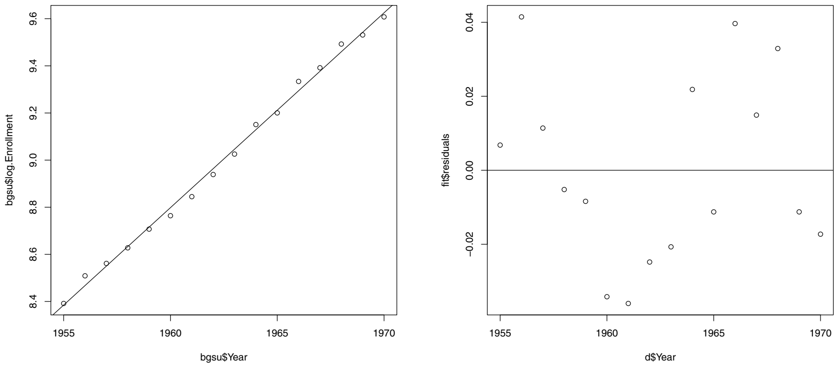
From the R output, we see that the least-squares fit to the log enrollment data is

$$\log Enrollment = -153.257 + 0.0827Year.$$

This is equivalent to the exponential fit

$$Enrollment = \exp(-153.257 + 0.0827Year) \propto (1.086)^{Year},$$

where $\propto$ means "is proportional to." We see that BGSU's enrollment was increasing approximately 8.6% a year during the period between 1955 and 1970.



(a) (b)

Fig. 5.9 Least-squares fit to the log enrollment data (a) and the residual plot (b). Since there is no strong curvature pattern in the residuals, a linear fit seems more appropriate for the log enrollment than for the enrollment.

5.6 Exploring Fraction Data

5.6.1 Stemplot

Example 5.3 (Ratings of colleges (continued)).

One measure of quality of a university is the percentage of incoming students who graduated in the top ten percent of their high school class. Suppose we focus our attention at the "Top Ten" percentages for the Tier 1 colleges. We first use the `subset` function to extract the Tier 1 schools and put them in a new data frame `college1`:

```
> college1 = subset(college, Tier==1)
```

A stemplot of the percentages can be produced using the `stem` function.

```
> stem(college1$Top.10)
   4 | 3
   5 | 589
   6 | 344468
   7 | 355599
   8 | 02445556777888
   9 | 0022333456667777889
  10 | 0
```

5.6.2 Transforming fraction data

Since the percentages are left-skewed with a cluster of values in the 90's, it is a little difficult to summarize and hard to distinguish the schools with high percentages of "top ten" students. This suggests that we might be able to improve the display by an appropriate reexpression. For percentage data or equivalently fraction data that are piled up at the extreme values of 0 and 1, Tukey suggests the use of several reexpressions. The *folded fraction* is defined by

$$ff = f - (1 - f).$$

This reexpression expands the scale from the interval (0, 1) to the interval $(-1, 1)$; a fraction $f = 0.5$ is a folded fraction of $ff = 0$. The *folded root* or *froot* is defined as

$$froot = \sqrt{f} - \sqrt{1 - f}$$

and the *folded log* or *flog* is defined as

$$flog = \log(f) - \log(1 - f).$$

Figure 5.10 displays the values of these reexpression for particular values of the fraction f. This figure was created using the following R code:

```
f = c(0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95)
ff = f - (1 - f)
froot = sqrt(2  * f) - sqrt(2 * (1 - f))
flog = 1.15 * log10(f) - 1.15 * log10(1 - f)
D = data.frame(f, ff, froot, flog)
matplot(t(as.matrix(D)), 1:4, type="l", lty=1, lwd=1,
  xlab="FRACTION", ylab="TRANSFORMATION",
  xlim=c(-1.8, 2), ylim=c(0.5, 4.3))
matplot(t(as.matrix(D[c(1, 4, 7), ])),
  1:4, type="l", lwd=3, lty=1, add=TRUE)
lines(D[c(1, 7), 1], c(1, 1), lwd=2)
lines(D[c(1, 7) ,2], 2 * c(1, 1), lwd=2)
lines(D[c(1, 7), 3], 3 * c(1, 1), lwd=2)
lines(D[c(1, 7), 4], 4 * c(1, 1), lwd=2)
text(c(1.8, 1.5, 1.3, 1.3, 0, 0.5 ,1),
  c(4, 3, 2, 1, 0.8, 0.8, 0.8),
  c("flog", "froot", "ff", "f", "f=.05", "f=.5", "f=.95"))
```

Figure 5.10 illustrates several desirable properties of these reexpressions. First, they are symmetric reexpressions in the sense that the ff or froot or flog of f will be the negative of the ff or froot or flog of $1 - f$. Also the froot and flog rexpressions have the effect of expanding the scale for fractions close to 0 or 1.

We compute these reexpressions of the Top Ten percentages. To avoid problems with computing logs at percentages of 0 and 100, a value of 0.5 is added to the percentages of Top Ten and "not Top Ten" before the flog is taken.

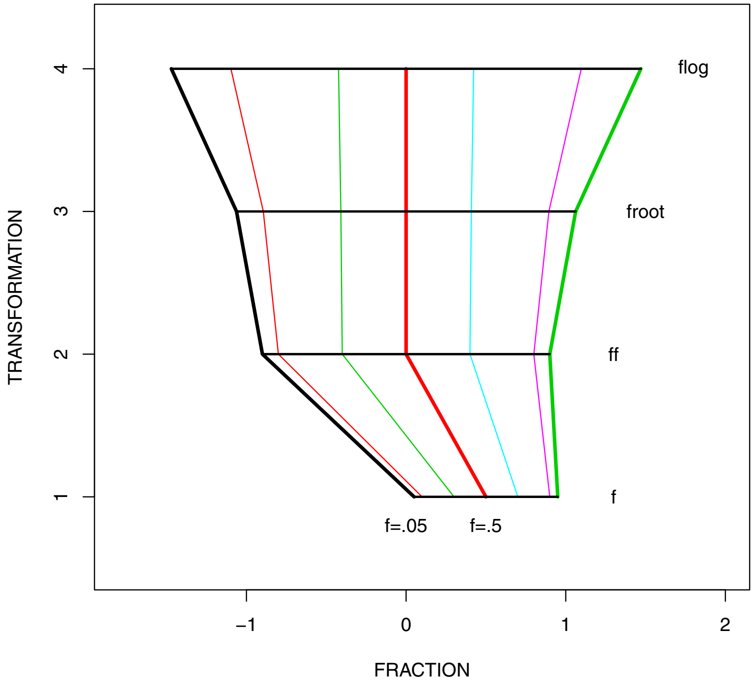


Fig. 5.10 Display of three different reexpressions for fraction data. The bottom line (labelled f) displays fraction values of 0.05, 0.10, 0.30, 0.50, 0.70, 0.90, 0.95, and the ff, froot, and flog lines display the fractions on the folded fraction, folded root, and folded log scales.

```
> froot = sqrt(college1$Top.10) - sqrt(100 - college1$Top.10)
> flog = log(college1$Top.10 + 0.5) - log(100 - college1$Top.10 + 0.5)
```

Stemplots of the froot and flog Top 10 percentages are displayed. Both reexpressions have the effect of making the Top 10 percentages more symmetric and spreading out the schools with high values.

```
> stem(froot}
  The decimal point is at the |

  -0 | 0
   0 | 7139
   2 | 000363777
   4 | 3358223335777999
   6 | 338800025888
   8 | 11111559
  10 | 0
```

```
> stem(flog)
  The decimal point is at the |

  -0 | 3
   0 | 234566677
   1 | 01113345667778999
   2 | 000224455579
   3 | 1113333377
   4 | 2
   5 | 3
```

What is the benefit of taking these strange-sounding reexpressions? On the froot scale, the percentages are approximately symmetric, and symmetric data has a clear "average." On the froot scale, a typical Top Ten percentage is 5.7. Also this reexpression can help to equalize spreads between groups, and provide a simple comparison. To illustrate, we use the subset function to create a data frame college34 with data from the Tier 3 and Tier 4 colleges. (The "|" symbol is the logical "or" operator; here we wish to include colleges that are either in Tier 3 or Tier 4.) We compute froots of the Top Ten percentages and use parallel boxplots to compare the two Tiers.

```
> college34 = subset(college, Tier==3 | Tier==4)
> froot = sqrt(college34$Top.10) - sqrt(100 - college34$Top.10)
> boxplot(froot ~ Tier, data=college34, horizontal=TRUE,
+   xlab="Froot(Top 10 Pct)", ylab="Tier")
```

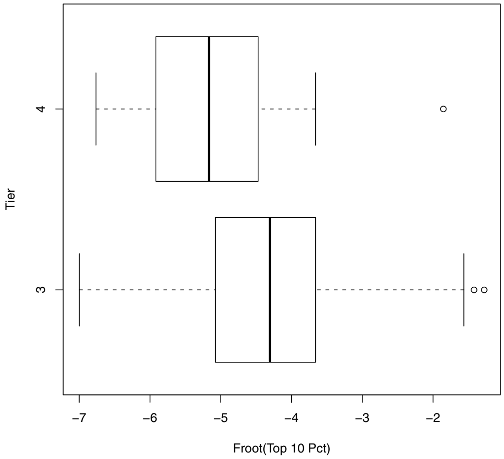


Fig. 5.11 Parallel boxplots of the froot Top 10 percentages of the Tier 3 and Tier 4 National universities.

We see from the display in Figure 5.11 that the Top 10 percentages, on the froot scale, have similar spreads that we measure by the quartile spread. One can compute the medians of the Top 10 froot percentages to be respectively -4.3 and -5.2. Therefore, on the froot scale, the Top Ten percentages for the Tier 3 schools tend to be $-4.3 - (-5.2) = 0.9$ higher than the Tier 4 schools.

Exercises

5.1 (Exploring percentages of small classes). The variable `Pct.20` in the college dataset contains the percentage of "small classes" (defined as 20 or fewer students) in the National Universities.

a. Construct a dotplot of the small-class percentages using the `stripchart` function. To see the density of points, it is helpful to use either the `method=stack` or `method=jitter` arguments. What is the shape of this data?
b. There is a single school with an unusually large small-class percentage. Use the `identify` function to find the name of this unusual school.
c. Find the median small-class percentage and draw a vertical line (using the `abline` function) on the dotplot at the location of the median.

5.2 (Relationship between the percentages of small classes and large classes). The variables `Pct.20` and `Pct.50` in the college dataset contain respectively the percentage of "small classes" (defined as 20 or fewer students) and the percentage of "large classes" (defined as 50 or more students) in the National Universities.

a. Use the `plot` function to construct a scatterplot of `Pct.20` (horizontal) against `Pct.50` (vertical).
b. Use the `line` function to find a resistant line to these data. Add this resistant line to the scatterplot constructed in part a.
c. If 60% of the classes at a particular college have 20 or fewer students, use the fitted line to predict the percentage of classes that have 50 or more students.
d. Construct a graph of the residuals (vertical) against `Pct.20` (horizontal) and add a horizontal line at zero (using the `abline` function).
e. Is there a distinctive pattern to the residuals? (Compare the sizes of the residuals for small `Pct.20` and the sizes of the residuals for large `Pct.50`.)
f. Use the `identify` function to identify the schools that have residuals that exceed 10 in absolute value. Interpret these large residuals in the context of the problem.

5.3 (Relationship between acceptance rate and "top-ten" percentage). The variables `Accept.rate` and `Top.10` in the college dataset contain respectively the acceptance rate and the percentage of incoming students in the top 10 percent of their high school class in the National Universities. One would believe that these two variables are strongly associated, since, for example, "exclusive" colleges with small acceptance rates would be expected to have a large percentage of "top-ten" students.

a. Explore the relationship between `Accept.rate` and `Top.10`. This exploration should include a graph and linear fit that describe the basic pattern in the relationship and a residual graph that shows how schools differ from the basic pattern.
b. Schools are often classified into "elite" and "non-elite" colleges depending on the type of students they admit. Based on your work in part a, is there any evidence from `Accept.rate` and `Top.10` that schools do indeed cluster into "elite" and "non-elite" groups? Explain.

5.4 (Exploring the pattern of college enrollment in the United States). The U.S. National Center for Education Statistics lists the total enrollment at Institutions of Higher Education for years 1900-1985 at their website `http://nces.ed.gov`. Define the ordered pair (x, y), where y is the total enrollment in thousands in year x. Then we observe the data (1955, 2653), (1956, 2918), (1957, 3324), (1959, 3640), (1961, 4145), (1963, 4780), (1964, 5280), (1965, 5921), (1966, 6390), (1967, 6912), (1968, 7513), (1969, 8005), (1970, 8581).

a. Enter this data into R.
b. Use the `lm` function to fit a line to the pattern of enrollment growth in the period 1955 to 1970. By inspecting a graph of the residuals, decide if a line is a reasonable model of the change in enrollment.
c. Transform the enrollment by a logarithm, and fit a line to the (year, log enrollment) data. Inspect the pattern of residuals and explain why a line is a better fit to the log enrollment data.
d. By interpreting the fit to the log enrollment data, explain how the college enrollment is changing in this time period. How does this growth compare to the growth of the BGSU enrollment in Section 5?

5.5 (Exploring percentages of full-time faculty). The variable `Full.time` in the college dataset (see Example 5.3) contains the percentage of faculty who are hired full-time in the group of National Universities.

a. Using the `hist` function, construct a histogram of the full-time percentages and comment on the shape of the distribution.
b. Use the froot and flog transformations to reexpress the full-time percentages. Construct histograms of the collection of froots and the collection of flogs. Is either transformation successful in making the full-time percentages approximately symmetric?
c. For data that is approximately normally distributed, about 68% of the data fall within one standard deviation of the mean. Assuming you have found a transformation in part (b) that makes the full-time percentages

approximately normal, find an interval that contains roughly 68% of the data on the new scale.

5.6 (Exploring alumni giving rates). The variable `Alumni.giving` contains the percentage of alumni from the college who make financial contributions.

a. Construct a "stacked" dotplot of the alumni giving percentages using the `stripchart` function.
b. Identify the names of the three schools with unusually large giving percentages.
c. It can be difficult to summarize these giving percentages since the distribution is right-skewed. One can make the dataset more symmetric by applying either a square root transformation or a log transformation.

```
roots = sqrt(college$Alumni.giving)
logs = log(college$Alumni.giving)
```

Apply both square root and log transformations. Which transformation makes the alumni giving rates approximately symmetric?

5.7 (Exploring alumni giving rates (continued)). In this exercise, we focus on the comparison of the alumni giving percentages between the four tiers of colleges.

a. Using the `stripchart` function with the stacked option, construct parallel dotplots of alumni giving by tier.
b. As one moves from Tier 4 to Tier 1, how does the average giving change?
c. As one moves from Tier 4 to Tier 1, how does the spread of the giving rates change?
d. We note from parts (b) and (c), that small giving rates tend to have small variation, and large giving rates tend to have large variation. One way of removing the dependence of average with spread is to apply a power transformation such as a square root or a log. Construct parallel stripcharts of the square roots of the giving rates, and parallel boxplots of the log giving rates.
e. Looking at the two sets of parallel stripcharts in part (d), were the square root rates or the log rates successful in making the spreads approximately the same between groups?