Table 3.1 Four Contrived Regression Data Sets From Anscombe (1973)

$X_{a,b,c}$	Y_a	Y_b	Y_c	X_d	Y_d
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

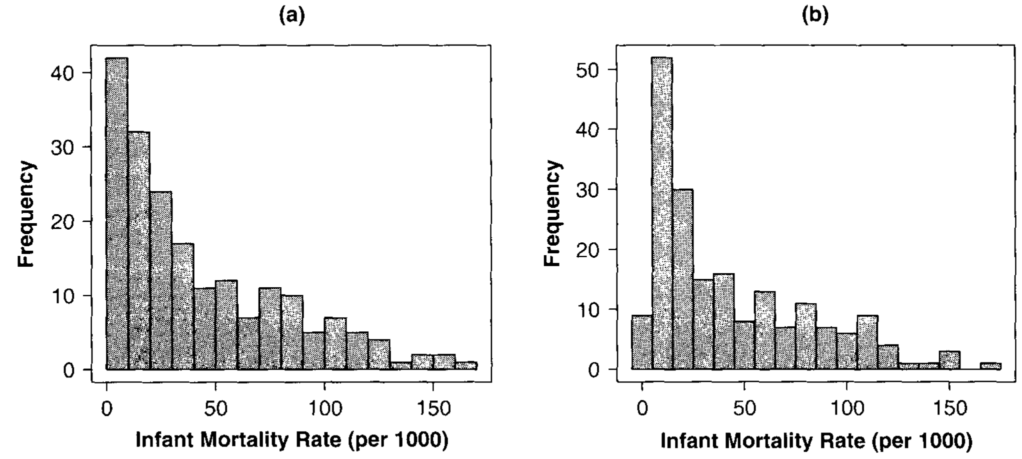


Figure 3.2 Histograms of infant morality for 193 nations. The histograms both use bins of width 10; histogram (a) employs bins that start at 0, while (b) employs bins that start at −5.

SOURCE: United Nations (1998).

3.1 Univariate Displays

3.1.1 Histograms

Figure 3.2 shows two *histograms* for the distribution of infant mortality among 193 countries, as reported in 1998 by the United Nations. The infant mortality rate is expressed as number of deaths of children aged less than 1 year per 1,000 live births. I assume that the histogram is a familiar graphical display, so I will offer only a brief description: To construct a histogram for infant mortality, dissect the range of the variable into equal-width intervals (called "bins"); count the number of observations falling in each bin; and display the frequency counts in a bar graph.

Both histograms in Figure 3.2 use bins of width 10; they differ in that the bins in Figure 3.2(a) start at 0 (i.e., 0 to 10, 10 to 20, etc.), while those in Figure 3.2(b) start at −5 (i.e., −5 to 5,

```
 1 | 2:
 leaf

 39
 72
 95
(19)
 79
 67
 56
 48
 39
 28
 22
 16
 10
  6
  5

HI: 153
```

Figure 3.3 Stem-and-le

5 to 15, etc.).[7] The two
for example, show that th
slightly different impress

Figure 3.3 shows an a
and-leaf plot, introduced
to form the bars of the h
display by hand to "scrat

You may be familia
explanation:

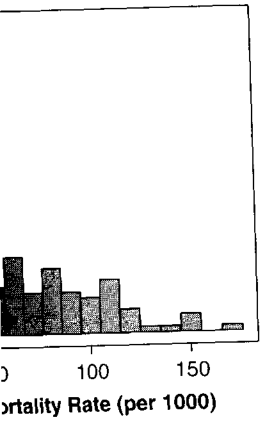
- Each data value is
Figure 3.3, the brea
mortality rate in Al
- Stems (here, 0, 1, .
bins, each of width
stem are then sortec
each stem into two
4–5, 6–7, 8–9); for
width 5 and five-pa
stems from leaves l
produce a display v
between the hundre
- Unusually large val
vidually. Here, ther
there countries witl
and displayed indiv
- The column of *dep*
The median is the
observations. For tl

[7] Because infant mortality can

[8] The rule for identifying outlie

Y_d
6.58
5.76
7.71
8.84
8.47
7.04
5.25
12.50
5.56
7.91
6.89

(b)



ns both use bins of width
loys bins that start at −5.

ality among 193 countries,
is expressed as number of
me that the histogram is a
construct a histogram for
rvals (called "bins"); count
ncy counts in a bar graph.
at the bins in Figure 3.2(a)
start at −5 (i.e., −5 to 5,

```
1 | 2: represents 12
 leaf unit: 1
         n: 193

   39      0 | 345555555666666667777777777778888999999
   72      1 | 000122222333344444555566778888999
   95      2 | 00112223333444455556669
  (19)     3 | 0001233445577889999
   79      4 | 012344456889
   67      5 | 11246667888
   56      6 | 01255568
   48      7 | 122347788
   39      8 | 00222456669
   28      9 | 025678
   22     10 | 234677
   16     11 | 023445
   10     12 | 2445
    6     13 | 2
    5     14 | 29

HI: 153 [Liberia], 154 [Afghanistan), 169 [Sierra Leone]
```

Figure 3.3 Stem-and-leaf display for infant mortality.

5 to 15, etc.).[7] The two histograms for infant mortality are more similar than different—both, for example, show that the distribution of infant mortality is positively skewed—but they do give slightly different impressions of the shape of the distribution.

Figure 3.3 shows an alternative form of histogram, called a *stem-and-leaf display*. The stem-and-leaf plot, introduced by John Tukey (1972, 1977), ingeniously employs the numerical data to form the bars of the histogram. As Tukey suggests, it is simple to construct a stem-and-leaf display by hand to "scratch down" a small data set.

You may be familiar with the stem-and-leaf display. Here is a relatively compressed explanation:

- Each data value is broken between two adjacent digits into a "stem" and a "leaf": In Figure 3.3, the break takes place between the tens and units digits. For example, the infant mortality rate in Albania was 32, which translates into the stem 3 and leaf 2.
- Stems (here, 0, 1, . . . , 14) are constructed to cover the data, implicitly defining a system of bins, each of width 10. Each leaf is placed to the right of its stem, and the leaves on each stem are then sorted into ascending order. We can produce a finer system of bins by dividing each stem into two parts (taking, respectively, leaves 0–4 and 5–9), or five parts (0–1, 2–3, 4–5, 6–7, 8–9); for the infant mortality data, two-part stems would correspond to bins of width 5 and five-part stems to bins of width 2. We could employ still finer bins by dividing stems from leaves between the ones and tenths digits, but, for infant mortality, that would produce a display with almost as many bins as observations. Similarly, a coarser division between the hundreds and tens digits would yield only two stems—0 and 1.
- Unusually large values—*outliers*—are collected on a special "HI" stem and displayed individually. Here, there are three countries with unusually large infant mortality rates. Were there countries with unusually small infant mortality rates, then these would be collected and displayed individually on a "LO" stem.[8]
- The column of *depths* counts in toward the median from both ends of the distribution. The median is the observation at depth $(n + 1)/2$, where (as usual) n is the number of observations. For the infant mortality data, the median is at depth $(193 + 1)/2 = 97$. In

[7] Because infant mortality cannot be negative, the contrast between Figures 3.2(a) and (b) is somewhat artificial.

[8] The rule for identifying outliers is explained in Section 3.1.4 on boxplots.

Figure 3.3, there are 39 observations at stem 0, 72 at and below stem 1, and so on; there are five observations (including the outliers) at and above stem 14, six at and above stem 13, and so forth. The count at the stem containing the median is shown in parentheses—here, 19 at stem 3. Note that 95 + 19 + 79 = 193.

In constructing histograms (including stem-and-leaf displays), we want enough bins to preserve some detail, but not so many that the display is too rough and dominated by sampling variation. Let n^* represent the number of nonoutlying observations. Then, for $n^* \leq 100$, it usually works well to use no more than about $2\sqrt{n^*}$ bins; likewise, for $n^* > 100$, we can use a maximum of about $10 \times \log_{10} n^*$ bins. Of course, in constructing a histogram, we also want bins that start and end at "nice" numbers (e.g., 10 to 20 rather than 9.5843 to 21.0457); in a stem-and-leaf display, we are limited to bins that correspond to breaks between digits of the data values. Computer programs that construct histograms incorporate rules such as these.[9]

For the distribution of infant mortality, $n^* = 193 - 3 = 190$, so we should aim for no more than $10 \times \log_{10}(190) \approx 23$ bins. The stem-and-leaf display in Figure 3.3 uses 15 stems (plus the "HI" stem).

Histograms, including stem-and-leaf displays, are very useful graphs, but they suffer from several problems:

- As we have seen, the visual impression of the data conveyed by a histogram can depend on the arbitrary origin of the bin system.
- Because the bin system dissects the range of the variable into class intervals, the histogram is discontinuous (i.e., rough) even if, as in the case of infant mortality, the variable is continuous.[10]
- The form of the histogram depends on the arbitrary width of the bins.
- Moreover, if we use bins that are narrow enough to capture detail where data are plentiful—usually near the center of the distribution—then they may be too narrow to avoid "noise" where data are sparse—usually in the tails of the distribution.

3.1.2 Nonparametric Density Estimation

Nonparametric density estimation addresses the deficiencies of traditional histograms by averaging and smoothing. As the term implies, "density estimation" can be construed formally as an attempt to estimate the probability density function of a variable based on a sample, but it can also be thought of informally as a descriptive technique for smoothing histograms.

In fact, the histogram—suitably rescaled—is a simple density estimator.[11] Imagine that the origin of the bin system is at x_0, and that each of the m bins has width $2h$; the end points of the

[9] More sophisticated rules for the number of bins take into account information beyond n. For example, Freedman and Diaconis (1981) suggest

$$\text{number of bins} \approx \left\lceil \frac{n^{1/3}\left(x_{(n)} - x_{(1)}\right)}{2(Q_3 - Q_1)} \right\rceil$$

where $x_{(n)} - x_{(1)}$ is the range of the data, $Q_3 - Q_1$ is the inter-quartile range, and the "ceiling" brackets indicate rounding *up* to the next integer.

[10] That is, infant mortality rates are continuous for practical purposes in that they can take on many different values. Actually, infant mortality rates are ratios of integers and hence are rational numbers, and the rates in the U.N. data set are rounded to the nearest whole number.

[11] Rescaling is required because a density function encloses a total area of 1. Histograms are typically scaled so that the height of each bar represents frequency (or percent), and thus the heights of the bars sum to the sample size n (or 100). If each bar spans a bin of width $2h$ (anticipating the notation below), then the total area enclosed by the bars is $n \times 2h$. Dividing the height of each bar by $2nh$ therefore produces the requisite density rescaling.

bins are then at x_0, x
convention)

The histogram estim
of observations that

where # is the count
We can dispense
continuously movin

In practice, of cours
values covering the r
equivalent to locally

where

a formulation that wi
the density. Here z i
$(x - X_i)/h$. The naiv
and is similar in spir
An illustration, u
principal problem wi
observations enter ar
The rectangular w
producing a density e
property—probabilit
called a *kernel*. Choo
edges of the naive de
The general kerne

There are many reas
normal density funct
in effect sums suitab

bins are then at $x_0, x_0 + 2h, x_0 + 4h, \ldots, x_0 + 2mh$. An observation X_i falls in the jth bin if (by convention)

$$x_0 + 2(j-1)h \le X_i < x_0 + 2jh$$

The histogram estimator of the density at any x value located in the jth bin is based on the number of observations that fall in that bin:

$$\widehat{p}(x) = \frac{\#_{i=1}^{n}[x_0 + 2(j-1)h \le X_i < x_0 + 2jh]}{2nh}$$

where # is the counting operator.

We can dispense with the arbitrary origin x_0 of the bin system by counting locally within a continuously moving window of half-width h centered at x:

$$\widehat{p}(x) = \frac{\#_{i=1}^{n}(x - h \le X_i < x + h)}{2nh}$$

In practice, of course, we would use a computer program to evaluate $\widehat{p}(x)$ at a large number of x values covering the range of X. This "naive density estimator" (so named by Silverman, 1986) is equivalent to locally weighted averaging, using a rectangular weight function:

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right) \tag{3.1}$$

where

$$W(z) = \begin{cases} \frac{1}{2} & \text{for } |z| < 1 \\ 0 & \text{otherwise} \end{cases}$$

a formulation that will be useful below when we consider alternative weight functions to smooth the density. Here z is a "stand-in" for the argument to the $W(\cdot)$ weight function—that is, $z = (x - X_i)/h$. The naive estimator is like a histogram that uses bins of width $2h$ but has no fixed origin, and is similar in spirit to the naive nonparametric-regression estimator introduced in Chapter 2.

An illustration, using the U.N. infant mortality data, appears in Figure 3.4, and reveals the principal problem with the naive estimator: Because the estimated density jumps up and down as observations enter and leave the window, the naive density estimator is intrinsically rough.

The rectangular weight function $W(z)$ in Equation 3.1 is defined to enclose an area of $2 \times \frac{1}{2} = 1$, producing a density estimate that (as required) also encloses an area of 1. Any function that has this property—probability density functions are obvious choices—may be used as a weight function, called a *kernel*. Choosing a kernel that is smooth, symmetric, and unimodal smooths out the rough edges of the naive density estimator. This is the essential insight of *kernel density estimation*.

The general kernel density estimator is, then, given by

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

There are many reasonable choices of the kernel function $K(z)$, including the familiar standard normal density function, $\phi(z)$, which is what I will use here. While the naive density estimator in effect sums suitably scaled rectangles centered at the observations, the more general kernel

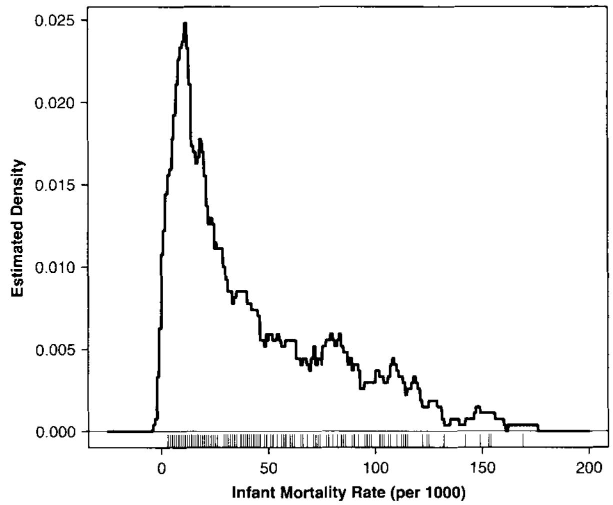


Figure 3.4 Naive density estimator for infant mortality, using a window half-width of $h = 7$. Note the roughness of the estimator. A *rug-plot* (or "one-dimensional scatterplot") appears at the bottom of the graph, showing the location of the data values.

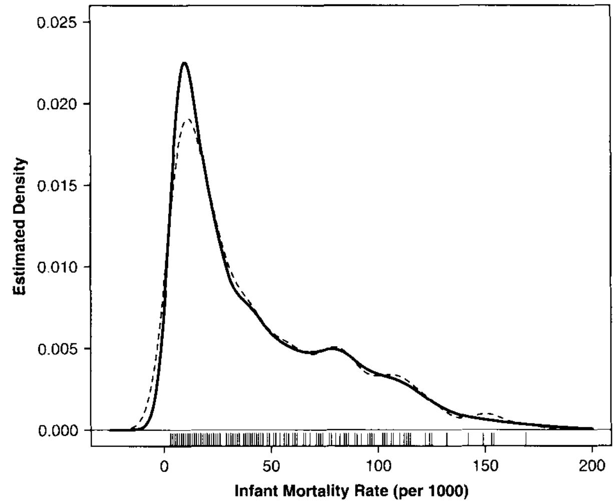


Figure 3.5 Kernel (broken line) and adaptive-kernel (solid line) density estimates for the distribution of infant mortality, using a normal kernel and a window half-width of $h = 7$. Note the relative "lumpiness" of the kernel estimator at the right, where data are sparse.

estimator sums smooth lumps. An example is shown in Figure 3.5, in which the kernel density estimator is given by the broken line.[12]

Selecting the window width for the kernel estimator is primarily a matter of trial and error—we want a value small enough to reveal detail but large enough to suppress random noise. We can,

[12] Notice that there is nonzero estimated density in Figure 3.5 below an infant mortality rate of 0. Of course, this does not make sense, and although I will not pursue it here, it is possible to constrain the lower and upper limits of the kernel estimator.

however, look to statistical trying to estimate is normal most efficient with the win

As is intuitively reasonable increased, permitting finer

Although we might, by sample standard deviation sufficiently non-normal, the compromise is to use an "a

The factor 1.349 is the interc range)/1.349 a robust estir

One further caveat: If th skewed or multimodal—the that is too wide. A good pr

and to adjust this value d is the procedure that was (interquartile range)/1.349 $h = 0.9 \times 38.55 \times 197^{-1/}$

The kernel density esti remains a compromise: We detail) and a wider one whe refer implicitly to the unde an initial estimate of the de estimate.[15] The result is t estimator of spread in Equa

1. Calculate an initial d
2. Using the initial estin at the observations:

In this formula, $\widetilde{p}$ is t that is,

[13] See, for example, Silverman (1

[14] If we really knew that the dens substituting the sample mean $\overline{X}$ $(2\pi\sigma^2)^{-1/2}\exp[-(x-\mu)^2/2\sigma^2$

[15] An alternative is to use a *ne* Chapter 2.

however, look to statistical theory for rough guidance:[13] If the underlying density that we are trying to estimate is normal with standard deviation σ, then (for the normal kernel) estimation is most efficient with the window half-width

$$h = 0.9\sigma n^{-1/5} \tag{3.2}$$

As is intuitively reasonable, the optimal window grows gradually narrower as the sample size is increased, permitting finer detail in large samples than in small ones.[14]

Although we might, by reflex, be tempted to replace the unknown σ in Equation 3.2 with the sample standard deviation S, it is prudent to be more cautious, for if the underlying density is sufficiently non-normal, then the sample standard deviation may be seriously inflated. A common compromise is to use an "adaptive" estimator of spread:

$$A = \min\left(S, \frac{\text{interquartile range}}{1.349}\right) \tag{3.3}$$

The factor 1.349 is the interquartile range of the standard normal distribution, making (interquartile range)/1.349 a robust estimator of σ in the normal setting.

One further caveat: If the underlying density is substantially non-normal—in particular, if it is skewed or multimodal—then basing h on the adaptive estimator A generally produces a window that is too wide. A good procedure, then, is to start with

$$h = 0.9An^{-1/5}$$

and to adjust this value downwards until the resulting density plot becomes too rough. This is the procedure that was used to find the window width in Figure 3.5, where $S = 38.55$ and (interquartile range)/1.349 = (68 − 13)/1.349 = 40.77. Here, the "optimal" window width is $h = 0.9 \times 38.55 \times 197^{-1/5} = 12.061$.

The kernel density estimator usually does a pretty good job, but the window half-width h remains a compromise: We would prefer a narrower window where data are plentiful (to preserve detail) and a wider one where data are sparse (to suppress noise). Because "plentiful" and "sparse" refer implicitly to the underlying density that we are trying to estimate, it is natural to begin with an initial estimate of the density, and to adjust the window half-width on the basis of the initial estimate.[15] The result is the *adaptive-kernel estimator* (not to be confused with the adaptive estimator of spread in Equation 3.3).

1. Calculate an initial density estimate, $\widetilde{p}(x)$—for example, by the kernel method.
2. Using the initial estimate, compute local window factors by evaluating the estimated density at the observations:

   $$f_i = \left[\frac{\widetilde{p}(X_i)}{\widetilde{p}}\right]^{-1/2}$$

   In this formula, $\widetilde{p}$ is the geometric mean of the initial density estimates at the observations—that is,

   $$\widetilde{p} = \left[\prod_{i=1}^{n} \widetilde{p}(X_i)\right]^{1/n}$$

[13] See, for example, Silverman (1986, chap. 3) for a detailed discussion of these issues.

[14] If we really knew that the density were normal, then it would be even more efficient to estimate it parametrically by substituting the sample mean $\overline{X}$ and standard deviation S for μ and σ in the formula for the normal density, $p(x) = (2\pi\sigma^2)^{-1/2}\exp[-(x-\mu)^2/2\sigma^2]$.

[15] An alternative is to use a *nearest-neighbor* approach, as in the nonparametric-regression methods discussed in Chapter 2.

(where the operator $\prod$ indicates continued multiplication). As a consequence of this definition, the f_is have a product of 1, and hence a geometric mean of 1, ensuring that the area under the density estimate remains equal to 1.

3. Calculate the adaptive-kernel density estimator using the local window factors to adjust the width of the kernels centered at the observations:

$$\widehat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{f_i} K\left(\frac{x - X_i}{f_i h}\right)$$

Applying the adaptive kernel estimator to the infant mortality distribution produces the solid line in Figure 3.5: For this distribution the kernel and adaptive-kernel estimates are very similar, although the adaptive kernel more sharply defines the principal mode of the distribution near 20, and produces a smoother long right tail.

3.1.3 Quantile-Comparison Plots

Quantile-comparison plots are useful for comparing an empirical sample distribution with a theoretical distribution, such as the normal distribution—something that is more commonly of interest for derived quantities such as test statistics or residuals than for observed variables. A strength of the display is that it does not require the use of arbitrary bins or windows.

Let $P(x)$ represent the theoretical *cumulative distribution function* (CDF) with which we want to compare the data; that is, $P(x) = \Pr(X \leq x)$. A simple (but not terribly useful) procedure is to graph the *empirical cumulative distribution function* (ECDF) for the observed data, which is simply the proportion of data below each value of x, as x moves continuously from left to right:

$$\widehat{P}(x) = \frac{\#_{i=1}^{n}(X_i \leq x)}{n}$$

As illustrated in Figure 3.6, however, the ECDF is a "stair-step" function (where each step occurs at an observation, and is of height $1/n$), while the CDF is typically smooth, making the comparison difficult.

The quantile-comparison plot avoids this problem by never constructing the ECDF explicitly:

1. Order the data values from smallest to largest, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The $X_{(i)}$ are called the *order statistics* of the sample.

2. By convention, the cumulative proportion of the data "below" $X_{(i)}$ is given by[16]

$$P_i = \frac{i - \frac{1}{2}}{n}$$

3. Use the inverse of the CDF (that is, the *quantile function*) to find the value z_i corresponding to the cumulative probability P_i; that is,[17]

$$z_i = P^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$$

[16] This definition avoids cumulative proportions of 0 or 1, which would be an embarrassment in step 3 for distributions, like the normal, that never quite reach cumulative probabilities of 0 or 1. In effect, we count half of each observation below its exact value and half above. Another common convention is to use $P_i = \left(i - \frac{1}{3}\right) / \left(n + \frac{1}{3}\right)$.

[17] This operation assumes that the CDF has an inverse—that is, that P is a strictly increasing function (one that never quite levels off). The common continuous probability distributions in statistics—for example, the normal, t-, F-, and χ^2 distributions—all have this property. These and other distributions are reviewed in Appendix D on probability and estimation.

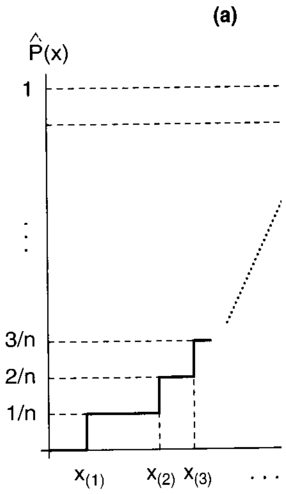


Figure 3.6 A "typical" empi
"typical" theoreti
represent the dat
values are not, in

4. Plot the z_i as horizonta
from the distribution
with an intercept of 0
sampling error (see pc
plot is approximately
are identical except fc
from 1, $X_{(i)} \approx \sigma z_i$; fir
same shape, then $X_{(i)}$

5. It is often helpful to
departures from linea
the data, or we can dra
plot—comparing the
alternatively use the
as a robust estimator
work well when the d

6. We expect some depa
interpretation to disp
error of the order stat

where $p(z)$ is the pr
ues along the fitted l
"envelope" around th

[18] By the method of construction,
There is a greater probability that

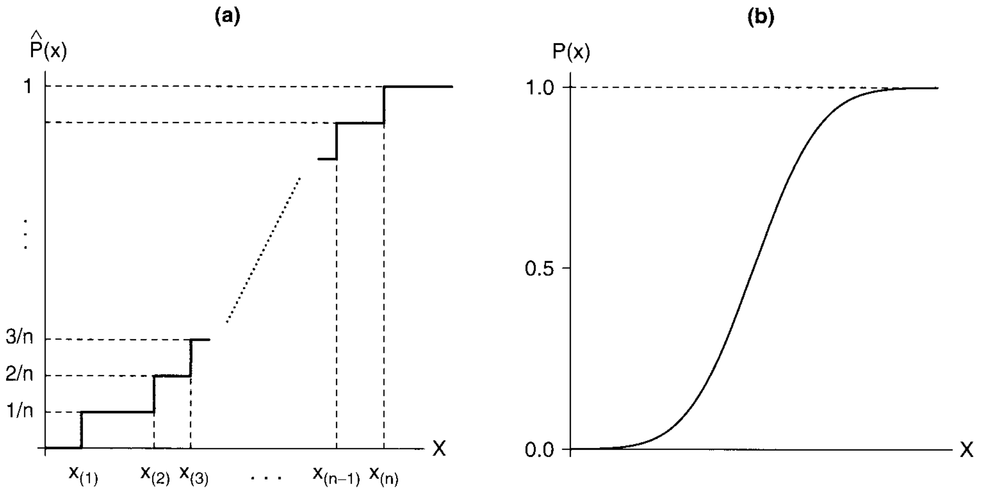


Figure 3.6 A "typical" empirical cumulative distribution function (ECDF) is shown in (a), a "typical" theoretical cumulative distribution function (CDF) in (b). $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represent the data values ordered from smallest to largest. Note that the ordered data values are not, in general, equally spaced.

4. Plot the z_i as horizontal coordinates against the $X_{(i)}$ as vertical coordinates. If X is sampled from the distribution P, then $X_{(i)} \approx z_i$. That is, the plot should be approximately linear, with an intercept of 0 and slope of 1. This relationship is only approximate because of sampling error (see point 6). If the distributions are identical except for location, then the plot is approximately linear with a nonzero intercept, $X_{(i)} \approx \mu + z_i$; if the distributions are identical except for scale, then the plot is approximately linear with a slope different from 1, $X_{(i)} \approx \sigma z_i$; finally, if the distributions differ both in location and scale but have the same shape, then $X_{(i)} \approx \mu + \sigma z_i$.

5. It is often helpful to place a comparison line on the plot to facilitate the perception of departures from linearity. The line can be plotted by eye, attending to the central part of the data, or we can draw a line connecting the quartiles. For a normal quantile-comparison plot—comparing the distribution of the data with the standard normal distribution—we can alternatively use the median as a robust estimator of μ and the interquartile range/1.349 as a robust estimator of σ. (The more conventional estimates $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$ will not work well when the data are substantially non-normal.)

6. We expect some departure from linearity because of sampling variation; it therefore assists interpretation to display the expected degree of sampling error in the plot. The standard error of the order statistic $X_{(i)}$ is

$$\text{SE}(X_{(i)}) = \frac{\widehat{\sigma}}{p(z_i)}\sqrt{\frac{P_i(1-P_i)}{n}}$$

where $p(z)$ is the probability density function corresponding to the CDF $P(z)$. The values along the fitted line are given by $\widehat{X}_{(i)} = \widehat{\mu} + \widehat{\sigma} z_i$. An approximate 95% confidence "envelope" around the fitted line is, therefore,[18]

[18] By the method of construction, the 95% confidence level applies (point-wise) to each $\widehat{X}_{(i)}$, not to the whole envelope: There is a greater probability that *at least one* point strays outside the envelope even if the data are sampled from the

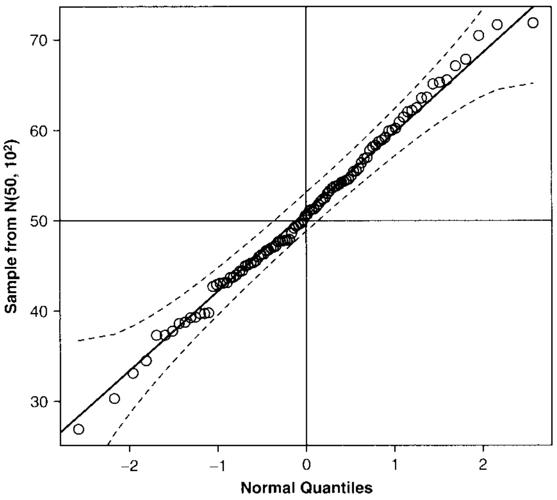


Figure 3.7 Normal quantile-comparison plot for a sample of 100 observations drawn from a normal distribution with mean 50 and standard deviation 10. The fitted line is through the quartiles of the distribution, and the broken lines give a point-wise 95% confidence interval around the fit.

$$\widehat{X}_{(i)} \pm 2 \times \text{SE}(X_{(i)})$$

Figures 3.7 to 3.10 display normal quantile-comparison plots for several illustrative distributions:

- Figure 3.7 plots a sample of $n = 100$ observations from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$. The plotted points are reasonably linear and stay within the rough 95% confidence envelope.
- Figure 3.8 plots a sample of $n = 100$ observations from the positively skewed chi-square distribution with 2 degrees of freedom. The positive skew of the data is reflected in points that lie *above* the comparison line in both tails of the distribution. (In contrast, the tails of negatively skewed data would lie *below* the comparison line.)
- Figure 3.9 plots a sample of $n = 100$ observations from the heavy-tailed t-distribution with 2 degrees of freedom. In this case, values in the upper tail lie above the corresponding normal quantiles, and values in the lower tail below the corresponding normal quantiles.
- Figure 3.10 shows the normal quantile-comparison plot for the distribution of infant mortality. The positive skew of the distribution is readily apparent. The possibly bimodal character of the data, however, is not easily discerned in this display.

Quantile-comparison plots highlight the tails of distributions. This is important, because the behavior of the tails is often problematic for standard estimation methods like least squares, but it is useful to supplement quantile-comparison plots with other displays—such as histograms or kernel-density estimates—that provide more intuitive representations of distributions. A key point is that there is no reason to limit ourselves to a single picture of a distribution when different pictures bring different aspects of the distribution into relief.

comparison distribution. Determining a *simultaneous* 95% confidence envelope would be a formidable task, because the order statistics are not independent.

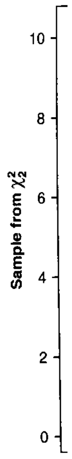


Figure 3.8 Normal quantile
positively skewe

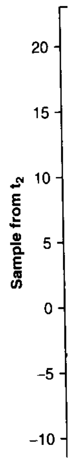


Figure 3.9 Normal quantil
heavy-tailed t-

3.1.4 Boxplots

Unlike histograms, dens
present only summary info
observations. Boxplots ar
minimum, first quartile, m
Boxplots, therefore, are us
example, in the margins of

vations drawn from a
). The fitted line is
s give a point-wise 95%

s for several illustrative

nal distribution with mean
reasonably linear and stay

sitively skewed chi-square
data is reflected in points
n. (In contrast, the tails of

heavy-tailed t-distribution
ie above the corresponding
onding normal quantiles.
istribution of infant mortal-
possibly bimodal character

s is important, because the
hods like least squares, but
plays—such as histograms
ons of distributions. A key
distribution when different

be a formidable task, because the

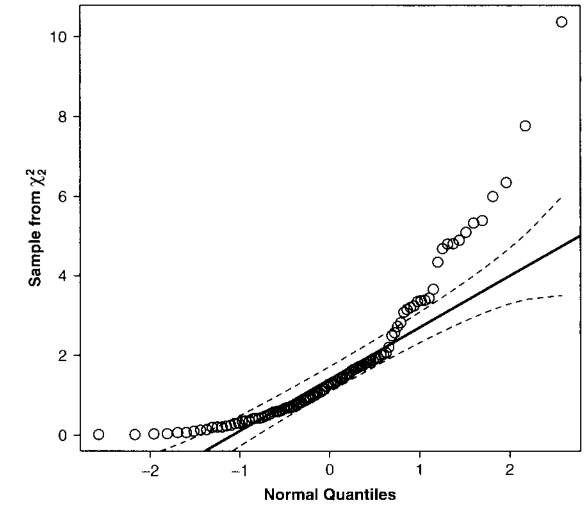


Figure 3.8 Normal quantile-comparison plot for a sample of 100 observations from the positively skewed chi-square distribution with 2 degrees of freedom.

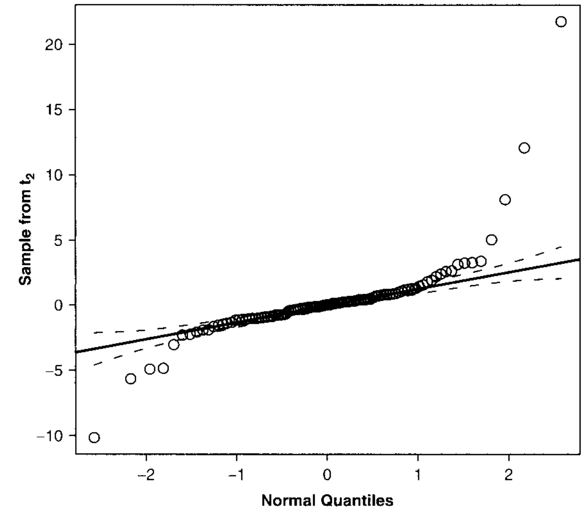


Figure 3.9 Normal quantile-comparison plot for a sample of 100 observations from the heavy-tailed t-distribution with 2 degrees of freedom.

3.1.4 Boxplots

Unlike histograms, density plots, and quantile-comparison plots, *boxplots* (due to Tukey, 1977) present only summary information on center, spread, and skewness, along with individual outlying observations. Boxplots are constructed from the *five-number summary* of a distribution—the minimum, first quartile, median, third quartile, and maximum—and outliers, if they are present. Boxplots, therefore, are useful when we require a compact representation of a distribution (as, for example, in the margins of a scatterplot), when we wish to compare the principal characteristics

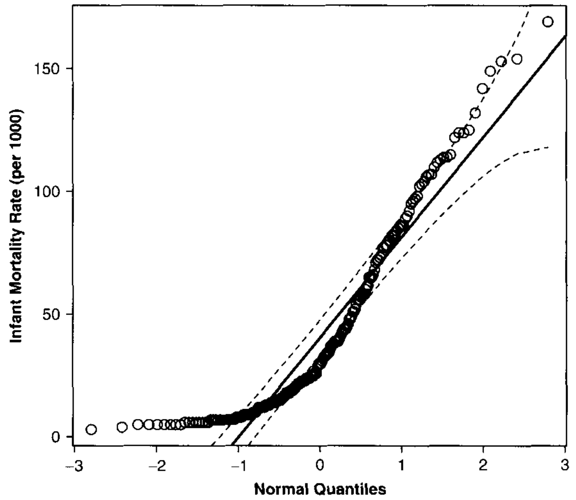


Figure 3.10 Normal quantile-comparison plot for the distribution of infant mortality. Note the positive skew.

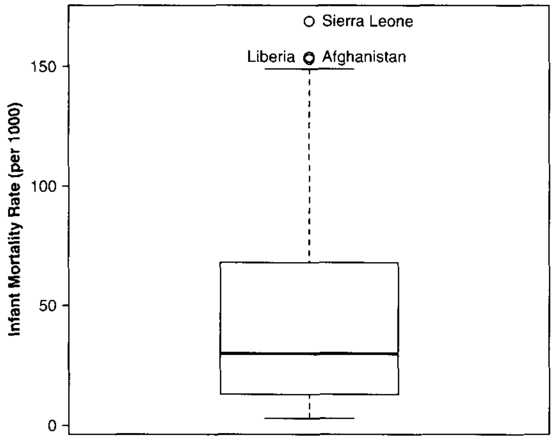


Figure 3.11 Boxplot for infant mortality. The central box is drawn between the hinges; the position of the median is marked in the box; and outlying observations are displayed individually.

of several distributions,[19] or when we want to select a transformation that makes a distribution more symmetric.[20]

An illustrative boxplot for infant mortality appears in Figure 3.11. This plot is constructed according to the following conventions (illustrated in the schematic horizontal boxplot in Figure 3.12):

1. A scale is laid off to accommodate the extremes of the data. The infant mortality data, for example, range between 3 and 169.

[19] See Section 3.2.

[20] Transformations to symmetry are discussed in Chapter 4.

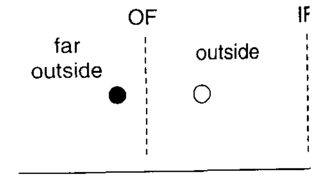


Figure 3.12 Schematic box
(adj), inner and

2. The central box is d
quartiles, and therefor
represents the median

giving the position o
largest: $X_{(1)}, X_{(2)}, \ldots$
part; using "floor" br
end to average the tw
depth(M) = (193 +
Likewise, the depth o

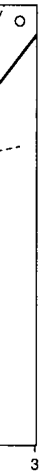
If depth(H) has a fra
the adjacent position
distribution, depth(H
and the upper hinge
yields the subscript 1

3. The following rules
boxplot:

- The *hinge-spread*

- The lower and up

Observations bey
termed "outside"
shown in the dis



ant mortality. Note the

een the hinges; the

bservations are displayed

that makes a distribution

. This plot is constructed

ic horizontal boxplot in

infant mortality data, for

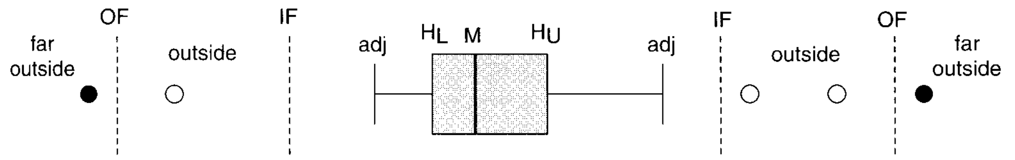


Figure 3.12 Schematic boxplot, showing the median (*M*), hinges (H_L and H_U), adjacent values (adj), inner and outer fences (IF and OF), and outside and far-outside observations.

2. The central box is drawn between the *hinges*, which are simply defined first and third quartiles, and therefore encompasses the middle half of the data. The line in the central box represents the median. Recall that the depth of the median is

$$\text{depth}(M) = \frac{n+1}{2}$$

giving the position of the middle observation after the data are ordered from smallest to largest: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. When n is even, the depth of the median has a fractional part; using "floor" brackets to represent truncation to an integer, we count in from either end to average the two observations at depth $\lfloor (n+1)/2 \rfloor$. For the infant mortality data, $\text{depth}(M) = (193+1)/2 = 97$, and $M = X_{(97)} = 30$.
Likewise, the depth of the hinges is

$$\text{depth}(H) = \frac{\lfloor \text{depth}(M) \rfloor + 1}{2}$$

If depth(H) has a fractional part, then, for each hinge, we average the two observations at the adjacent positions, that is, at $\lfloor \text{depth}(H) \rfloor$ and $\lfloor \text{depth}(H) \rfloor + 1$. For the infant mortality distribution, $\text{depth}(H) = (97+1)/2 = 49$. The lower hinge is, therefore, $H_L = X_{(49)} = 13$, and the upper hinge is $H_U = X_{(149)} = 73$. (Counting down 97 observations from the top yields the subscript $197 - 49 + 1 = 149$.)

3. The following rules are used to identify outliers, which are shown individually in the boxplot:

- The *hinge-spread* (roughly the interquartile range) is the difference between the hinges:

$$H\text{-spread} = H_U - H_L$$

- The lower and upper "inner fences" are located 1.5 hinge-spreads beyond the hinges:

$$\begin{aligned} \text{IF}_L &= H_L - 1.5 \times H\text{-spread} \\ \text{IF}_U &= H_U + 1.5 \times H\text{-spread} \end{aligned}$$

Observations beyond the inner fences (but within the outer fences, defined below) are termed "outside" and are represented by open circles. The fences themselves are not shown in the display.

- The "outer fences" are located three hinge-spreads beyond the hinges:[21]

$$\begin{aligned} \mathrm{OF}_L &= H_L - 3 \times H\text{-spread} \\ \mathrm{OF}_U &= H_U + 3 \times H\text{-spread} \end{aligned}$$

  Observations beyond the outer fences are termed "far outside" and are represented by filled circles. There are no far-outside observations in the infant mortality data.
- The "whisker" growing from each end of the central box extends either to the extreme observation on its side of the distribution (as at the low end of the infant mortality data) or to the most extreme nonoutlying observation, called the "adjacent value" (as at the high end of the infant mortality distribution).[22]

The boxplot of infant mortality in Figure 3.11 clearly reveals the skewness of the distribution: The lower whisker is much shorter than the upper whisker; the median is closer to the lower hinge than to the upper hinge; and there are several outside observations at the upper end of the infant mortality distribution, but none at the lower end. The apparent bimodality of the infant mortality data is not captured by the boxplot, however.

> **There are many useful univariate displays, including the traditional histogram. The stem-and-leaf plot is a modern variant of the histogram for small data sets, constructed directly from numerical data. Nonparametric density estimation may be employed to smooth a histogram. Quantile comparison plots are useful for comparing data with a theoretical probability distribution. Boxplots summarize some of the most important characteristics of a distribution, including center, spread, skewness, and outliers.**

3.2 Plotting Bivariate Data

The *scatterplot*—a direct geometric representation of observations on two quantitative variables (generically, Y and X)—is the most useful of all statistical graphs. The scatterplot is a natural representation of data partly because the media on which we draw plots—paper, computer screens—are intrinsically two dimensional. Scatterplots are as familiar and essentially simple as they are useful; I will therefore limit this presentation to a few points. There are many examples of bivariate scatterplots in this book, including in the preceding chapter.

- In analyzing data, it is convenient to work in a computing environment that permits the interactive identification of observations in a scatterplot.
- Because relationships between variables in the social sciences are often weak, scatterplots can be dominated visually by "noise." It often helps, therefore, to plot a nonparametric regression of Y on X.

[21] Here is a rough justification for the fences: In a normal population, the hinge-spread is 1.349 standard deviations, and so $1.5 \times H\text{-spread} = 1.5 \times 1.349 \times \sigma \approx 2\sigma$. The hinges are located $1.349/2 \approx 0.7$ standard deviations above and below the mean. The inner fences are, therefore, approximately at $\mu \pm 2.7\sigma$, and the outer fences at $\mu \pm 4.7\sigma$. From the standard normal table, $\Pr(Z > 2.7) \approx .003$, so we expect slightly less than 1% of the observations beyond the inner fences ($2 \times .003 = .006$); likewise, because $\Pr(Z > 4.7) \approx 1.3 \times 10^{-6}$, we expect less than one observation in 100,000 beyond the outer fences.

[22] All of the folksy terminology—"hinges," "fences," "whiskers," and so on—originates with Tukey (1977).

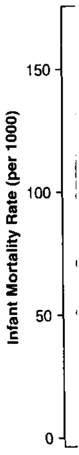


Figure 3.13 Scatterplot for
lowess smooth
levels of GDP

- Scatterplots in which
because the bulk of th
ple, the scatterplot fo
Figure 3.13. It often h
between Y and X.[23]
- Scatterplots in which
instance of this phen
vocabulary test agains
Surveys conducted b
and include in total 2
discrete—is to focus
for example, can be
discussed below). A
random quantity to th
random variable on t
Paradoxically, the ter
randomly "jittered"

> **The bivariate scatterpl
> quantitative variables.
> nonparametric regressi
> Scatterplots of the rela
> jittering the data.**

[23] See Chapter 4.

[24] The idea of jittering a scatterp