delivery suite, where Y_j is a gamma random variable with mean μ and variance σ^2. The values of these parameters are $\theta \doteq 13$, $\mu \doteq 8$ hours and $\sigma^2 \doteq 13$ hours squared. The average time and median times spent, $\overline{Y} = N^{-1}\sum Y_j$ and M, vary from day to day, with the lower right panel of Figure 2.1 suggesting that $\mathrm{E}(M) < \mathrm{E}(\overline{Y})$ and $\mathrm{var}(M) > \mathrm{var}(\overline{Y})$, properties we shall see theoretically in Example 2.30. ■

Much of this book is implicitly or explicitly concerned with distinguishing random and systematic variation. The notions of sampling variation and of a random sample are central, and before continuing we describe a useful tool for comparison of data and a distribution.

2.1.4 Probability plots

It is often useful to be able to check graphically whether data $y_1, \ldots, y_n$ come from a particular distribution. Suppose that in addition to the data we had a random sample $x_1, \ldots, x_n$ known to be from F. In order to compare the shapes of the samples, we could sort them to get $y_{(1)} \leq \cdots \leq y_{(n)}$ and $x_{(1)} \leq \cdots \leq x_{(n)}$, and make a *quantile-quantile* or *Q-Q plot* of $y_{(1)}$ against $x_{(1)}$, $y_{(2)}$ against $x_{(2)}$, and so forth. A straight line would mean that $y_{(j)} = a + bx_{(j)}$, so that the shape of the samples was identical, while distinct curvature would indicate systematic differences between them. If the line was close to straight, we could be fairly confident that $y_1, \ldots, y_n$ looks like a sample from F — after all, it would have a shape similar to the sample $x_1, \ldots, x_n$ which is from F.

Quantile-quantile plots are helpful for comparison of two samples, but when comparing a single sample with a theoretical distribution it is preferable to use F directly in a *probability plot*, in which the $y_{(j)}$ are graphed against the *plotting positions* $F^{-1}\{j/(n+1)\}$. This use of the $j/(n+1)$ quantile of F is justified in Section 2.3 as an approximation to $\mathrm{E}(X_{(j)})$, where $X_{(j)}$ is the random variable of which $x_{(j)}$ is a particular value. For example, the jth plotting positions for the normal and exponential distributions $\Phi\{(x-\mu)/\sigma\}$ and $1 - e^{-\lambda x}$ are $\mu + \sigma\Phi^{-1}\{j/(n+1)\}$ and $-\lambda^{-1}\log\{1 - j/(n+1)\}$. When parameters such as μ, σ, and λ are unknown, the plotting positions used are for standardized distributions, here $\Phi^{-1}\{j/(n+1)\}$ and $-\log\{1 - j/(n+1)\}$, which are sometimes called *normal scores* and *exponential scores*. Probability plots for the normal distribution are particularly common in applications and are also called *normal scores plots*. The interpretation of a probability plot is aided by adding the straight line that corresponds to perfect fit of F.

Example 2.16 (Birth data) The top left panel of Figure 2.3 shows a probability plot to compare the 95 times in the delivery suite with the normal distribution. The distribution does not fit the largest and smallest observations, and the data show some upward curvature relative to the straight line.

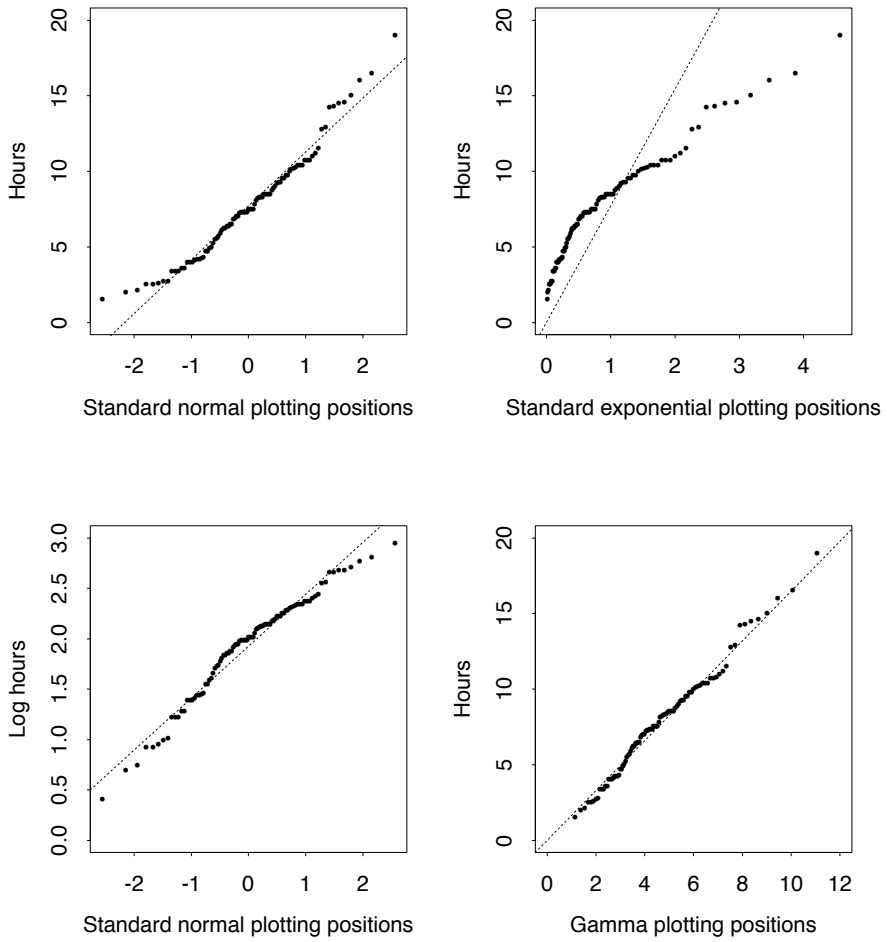

Figure 2.3
Probability plots for hours in the delivery suite, for the normal, exponential, gamma, and log-normal distributions (clockwise from top left). In each panel the dotted line is for a fitted distribution whose mean and variance match those of the data. None of the fits is perfect, but the gamma distribution fits best, and the exponential worst.

The top right panel shows that the exponential distribution would fit the data very poorly. The bottom left panel, a probability plot of the $\log y_j$ against normal plotting positions, corresponding to checking the log-normal distribution, shows slight downward curvature. The bottom right panel, a probability plot of the y_j against plotting positions for the gamma distribution with mean $\overline{y}$ and variance s^2, shows the best fit overall, though it is not perfect.

In the normal and gamma plots the dotted line corresponds to the theoretical distribution whose mean equals $\overline{y}$ and whose variance equals s^2; the dotted line in the exponential plot is for the exponential distribution whose mean equals $\overline{y}$; and the dotted line in the log-normal plot is for the normal

distribution whose mean and variance equal the average and variance of the $\log y_j$. ■

Some experience with interpreting probability plots may be gained from Practical 2.3.

Exercises 2.1

1 Let m and s be the values of location and scale statistics calculated from $y_1, \ldots, y_n$; m and s may be any of the quantities described in Examples 2.1 and 2.2. Show that the effect of the mapping $y_1, \ldots, y_n \mapsto a + by_1, \ldots, a + by_n$ $b > 0$, is to send $m, s \mapsto a + bm, bs$. Show also that the measures of shape in Examples 2.4 and 2.5 are unchanged by this transformation.

2 (a) Show that when δ is added to one of $y_1, \ldots, y_n$ and $|\delta| \to \infty$, the average $\overline{y}$ changes by an arbitrarily large amount, but the sample median does not. By considering such perturbations when n is large, deduce that the sample median has breakdown point 0.5. A sketch may help.
(b) Find the breakdown points of the other statistics in Examples 2.1 and 2.2.

3 (a) If $\kappa > 0$ is real and k a positive integer, show that the gamma function

$$\Gamma(\kappa) = \int_0^\infty u^{\kappa-1} e^{-u}\, du,$$

has properties $\Gamma(1) = 1$, $\Gamma(\kappa + 1) = \kappa\Gamma(\kappa)$ and $\Gamma(k) = (k-1)!$. It is useful to know that $\Gamma(\frac{1}{2}) = \pi^{1/2}$, but you need not prove this.
(b) Use (a) to verify the mean and variance of (2.7).
(c) Show that for $0 < \kappa \leq 1$ the maximum value of (2.7) is at $y = 0$, and find its mode when $\kappa > 1$. The *mode* of a density f is a value y such that $f(y) \geq f(x)$ for all x.

4 Give formulae analogous to (2.4) for the variance, skewness and 'shape' of a distribution F. Do they behave sensibly when a variable Y with distribution F is transformed to $a + bY$, so $F(y)$ is replaced by $F\{(y-a)/b\}$?

5 Let Y have continuous distribution function F. For any η, show that $X = |Y - \eta|$ has distribution $G(x) = F(\eta + x) - F(\eta - x)$, $x > 0$. Hence give a definition of the median absolute deviation of F in terms of F^{-1} and G^{-1}. If the density of Y is symmetric about the origin, show that $G(x) = 2F(x) - 1$. Hence find the median absolute deviation of the Laplace density (2.5).

6 A probability plot in which $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ are two random samples is called a *quantile-quantile* or *Q-Q plot*. Construct this plot for the first two columns in Table 2.1. Are the samples the same shape?

7 The *stem-and-leaf display* for the data 2.1, 2.3, 4.5, 3.3, 3.7, 1.2 is

```
1 | 2
2 | 13
3 | 37
4 | 5
```

If you turn the page on its side this gives a histogram showing the data values themselves (perhaps rounded); the units corresponding to intervals $[1, 2)$, $[2, 3)$ and so forth are to the left of the vertical bars, and the digits are to the right.