8
The Gaussian probability model

Until now we have been concerned only with the binary probability model. In this model there are two possible outcomes and the total probability of 1 is shared between them. It is an appropriate model when studying the occurrence of events, but not when studying a response for which there are many possible outcomes, such as blood pressure. For this the *Gaussian* or *normal* probability model is most commonly used.

In the Gaussian model the total probability of 1 is shared between many values. This is illustrated in the left panel of Fig. 8.1. When measurements are recorded to a fixed number of decimal places, there is a finite number of possible outcomes but, in principle, such measurements have infinitely many possible outcomes, so the probability attached to any one is effectively zero. For this reason it is the probability *density* per unit value which is specified by the model, not the probability of a given value. This is illustrated in the right panel of the figure. If π is the probability shared between values in a very narrow range, width h units, the probability density is π/h.

8.1 The standard Gaussian distribution

The standard Gaussian distribution has probability density centred at 0. The probability density at any value z (positive or negative) is given by

$$0.3989 \exp\left[-\frac{1}{2}(z)^2\right].$$

A graph of this probability density for different values of z is shown in Fig. 8.2. There is very little probability outside the range ± 3.

Tables of the standard Gaussian distribution are widely available, and these readily allow calculation of the probability associated with specified ranges of z. For our purposes it is necessary only to record that the probability corresponding to the range $(-1.645, +1.645)$ is 0.90 and that for the range $(-1.960, +1.960)$ is 0.95.

If the probability model for z is a standard Gaussian distribution then the probability model for $(z)^2$ is called the *chi-squared* distribution on one degree of freedom. Tables of chi-squared distributions can be used to find

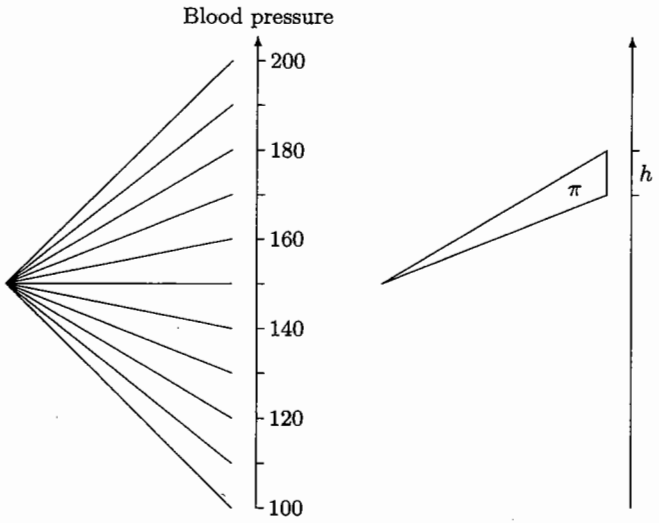

Fig. 8.1. Probability shared between many outcomes.

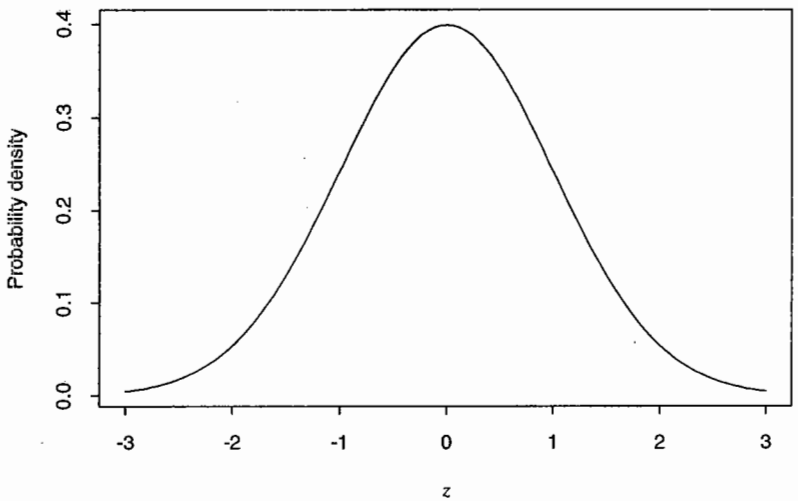

Fig. 8.2. The standard Gaussian distribution.

the probabilities of exceeding specified values of $(z)^2$ in the same way as tables of the standard Gaussian distribution are used to find probabilities of exceeding specified values of z.

Exercise 8.1. Use the tables in Appendix D to find the probability of exceeding the value 2.706 in a chi-squared distribution on one degree of freedom.

Note that, for $(z)^2$ to exceed 2.706, z must lie outside the range ± 1.645 of the standard normal distribution.

8.2 The general Gaussian model

It would be remarkable if the data we are analysing fell into the range -3 to $+3$, so for modelling the variability of real data, it is necessary to generalize the model to incorporate two parameters, one for the central value or *location*, and one for the spread or *scale* of the distribution. These are called the *mean* parameter and *standard deviation* parameter and are usually denoted by μ and σ respectively. A variable with such a distribution is derived by multiplying z by the scale factor and adding the location parameter. Thus

$$x = \mu + \sigma z.$$

has a distribution of the same general shape as the standard Gaussian distribution but centred around μ with most of its probability between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Exercise 8.2. If the mean and standard deviation of a general Gaussian distribution are 100 and 20 respectively, what ranges of values correspond to probabilities of 0.90 and 0.95 respectively?

Similarly, when x has a Gaussian distribution with mean μ and standard deviation σ then

$$z = \left(\frac{x-\mu}{\sigma}\right)$$

will have a *standard* Gaussian distribution. This fact can be used get the probability for a range of values of x using tables of z.

The probability density per unit of x when x has a Gaussian distribution with mean μ and standard deviation σ is

$$\frac{0.3989}{\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].$$

This expression is obtained by substituting $(x-\mu)/\sigma$ for z in the probability density of a standard Gaussian distribution to obtain the probability density per σ units of x, and then dividing by σ to obtain the probability density per unit of x. Sometimes the distribution is described in terms of the square of σ, which is called the *variance*.

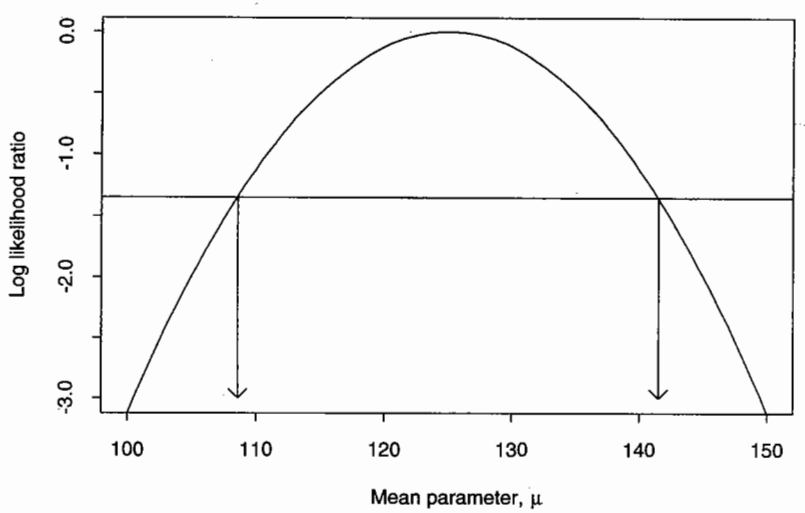

Fig. 8.3. The log likelihood ratio for the Gaussian mean, μ.

8.3 The Gaussian likelihood

Suppose a single value of x, say $x = 125$ is observed. Using the probability model that this is an observation from a Gaussian distribution with parameters μ and σ, the log likelihood for μ and σ is given by the log of the corresponding Gaussian probability density:

$$\log(0.3989) - \log(\sigma) - \frac{1}{2}\left(\frac{125-\mu}{\sigma}\right)^2.$$

This log likelihood depends on two unknown parameters, but to keep things simple we shall assume that one of them, σ, is known from past experience to have the value 10. Omitting constant terms, the log likelihood for μ is then

$$-\frac{1}{2}\left(\frac{125-\mu}{10}\right)^2.$$

The most likely value of μ is 125 and, since the above expression is zero at this point, this expression also gives the log likelihood *ratio* for μ. This is plotted in Fig. 8.3; curves with this shape are called *quadratic*.

We saw in Chapter 3 that we take the extremes of the supported range for a parameter to correspond to the value -1.353 for the log likelihood ratio. To find the limits of the supported range for μ we must therefore

solve the simple equation

$$-\frac{1}{2}\left(\frac{125-\mu}{10}\right)^2 = -1.353.$$

This takes only a few lines:

$$\begin{aligned}\left(\frac{125-\mu}{10}\right)^2 &= 2.706,\\ \left(\frac{125-\mu}{10}\right) &= \pm 1.645,\\ \mu &= 125 \pm 1.645 \times 10,\end{aligned}$$

so that supported values of μ are those between 108.6 and 141.5. In general, the log likelihood ratio for μ is

$$-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2,$$

the most likely value of μ is the observation x, and the supported range for μ is

$$x \pm 1.645\sigma,$$

where σ is the standard deviation (which we assume to be known).

We saw in Exercise 8.1 that the probability of exceeding 2.706 in a chi-squared distribution is 0.10, and the probability corresponding to the range ± 1.645 in the standard Gaussian distribution is 0.90. The fact that these numbers turn up in the above calculation is no accident and suggests that the log likelihood ratio criterion of -1.353 leads to supported ranges which have something to do with a probability of 0.90. This is indeed the case, but the relationship is not altogether straightforward and we shall defer this discussion to Chapter 10.

8.4 The likelihood with N observations

When there are N observations

$$x_1, x_2, \ldots, x_N,$$

the log likelihood for μ is obtained by adding the separate log likelihoods for each observation giving

$$\sum -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2.$$

Let M refer to the mean of the observations,

$$M = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

It can be shown that the log likelihood can be rearranged as

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2 + \sum -\frac{1}{2}\left(\frac{x_i - M}{\sigma}\right)^2$$

where $S = \sigma/\sqrt{N}$, sometimes called the *standard error of the mean.* This rearrangement involves only elementary algebra and the details are omitted. The second part of this new expression for the log likelihood does not depend on μ and cancels in the log likelihood ratio for μ which is

$$-\frac{1}{2}\left(\frac{M-\mu}{S}\right)^2,$$

The most likely value of μ is M, and setting the log likelihood ratio equal to -1.353 to obtain a supported range for μ gives

$$\mu = M \pm 1.645S.$$

As we would expect, with larger N, the value of S becomes smaller and the supported range narrower.

Exercise 8.3. The following measurements of systolic blood pressure were obtained from a sample of 20 men.

98	160	136	128	130	114	123	134	128	107
123	125	129	132	154	115	126	132	136	130

What is the most likely value for μ? Assuming that $\sigma = 14$, calculate the range of supported values for μ.

This exercise continues to make the unrealistic assumption, made throughout this chapter, that σ is *known.* In practice it must almost invariably be estimated from the data. We shall defer discussion of this until Chapter 34.

Solutions to the exercises

8.1 The probability of exceeding 2.706 in the chi-squared distribution with one degree of freedom is 0.10.

8.2 The range corresponding to a probability of 0.9 is

$$100 \pm 1.645 \times 20 = (67.1, 132.9)$$

and, for a probability of 0.95,

$$100 \pm 1.96 \times 20 = (60.8, 139.2).$$

8.3 The mean of the 20 measurements is 128.00 and this is the most likely value of μ. To calculate the supported range for μ, we first calculate

$$S = \frac{\sigma}{\sqrt{N}} = \frac{14}{\sqrt{20}} = 3.13$$

so that the range lies between

$$\mu = 128.00 \pm 1.645 \times 3.13$$

that is from 122.9 to 133.1 .