# 5

# INFERENCES FOR NORMAL DISTRIBUTIONS

In this chapter we begin the discussion of the topic that will occupy the rest of the book: the problem of inference, or how degrees of belief are altered by data. We start with the situation where the random variables that form the data have normal distributions. The reader may like to re-read §1.6, excluding the part that deals with the justification of the axioms, before starting the present chapter.

## 5.1. Bayes's theorem and the normal distribution

A *random sample of size n* from a distribution is defined as a set of $n$ independent random variables each of which has this distribution (cf. §§1.3, 3.3). If for each real number, $\theta$, belonging to a set (say, the set of positive numbers or the set of all real numbers), $f(x|\theta)$ is the density of a random variable, then $\theta$ is called a *parameter* of the *family* of distributions defined by the densities $\{f(x|\theta\}$ (cf. the parameter, $p$, of the binomial distribution, §2.1). We consider taking a random sample from a distribution with density $f(x|\theta)$ where $\theta$ is fixed but unknown and the function $f$ is known. Let $H$ denote our state of knowledge before the sample is taken. Then $\theta$ will have a distribution dependent on $H$; this will be a distribution of probability in the sense of degree of belief, and we denote its density by $\pi(\theta|H)$. As far as possible $\pi$ will be used for a density of beliefs, $p$ will be used for a density in the frequency sense, the sense that has been used in applications in chapters 2–4. If the random sample is $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ then the density of it will be, because the $x_i$ are independent,

$$\prod_{i=1}^{n} f(x_i|\theta) = p(\mathbf{x}|\theta, H), \quad \text{say.} \tag{1}$$

(The symbol $H$ should strictly also appear after $\theta$ on the left-hand side.) The density of beliefs about $\theta$ will be changed by

the sample according to Bayes's theorem (theorem 1.4.6 and its generalization, equation 3.2.9) into $\pi(\theta|\mathbf{x}, H)$ given by

$$\pi(\theta|\mathbf{x}, H) \propto p(\mathbf{x}|\theta, H)\,\pi(\theta|H) \tag{2}$$

according to the density form of the theorem (equation 3.2.9). The constant of proportionality omitted from (2) is

$$\left\{\int p(\mathbf{x}|\theta, H)\,\pi(\theta|H)\,d\theta\right\}^{-1} = \left\{\pi(\mathbf{x}|H)\right\}^{-1} \tag{3}$$

say, and does not involve $\theta$. $H$ will often be omitted from these and similar equations in agreement with the convention that an event which is always part of the conditioning event is omitted (§1.2). It will accord with the nomenclature of §1.6 if $\pi(\theta|H)$ is called the *prior* density of $\theta$; $p(\mathbf{x}|\theta, H)$, as a function of $\theta$, is called the *likelihood*; and $\pi(\theta|\mathbf{x}, H)$ is called the *posterior* density of $\theta$. We first consider the case of a single observation where $\mathbf{x} = x$ and $f(x|\theta)$ is the normal density.

***Theorem 1.*** *Let $x$ be $N(\theta, \sigma^2)$, where $\sigma^2$ is known, and the prior density of $\theta$ be $N(\mu_0, \sigma_0^2)$. Then the posterior density of $\theta$ is $N(\mu_1, \sigma_1^2)$, where*

$$\mu_1 = \frac{x/\sigma^2 + \mu_0/\sigma_0^2}{1/\sigma^2 + 1/\sigma_0^2}, \quad \sigma_1^{-2} = \sigma^{-2} + \sigma_0^{-2}. \tag{4}$$

(Effectively this is a result for a random sample of size one from $N(\theta, \sigma^2)$.) The likelihood is (omitting $H$)

$$p(x|\theta) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-(x-\theta)^2/2\sigma^2\right] \tag{5}$$

and the prior density is

$$\pi(\theta) = (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp\left[-(\theta-\mu_0)^2/2\sigma_0^2\right] \tag{6}$$

so that, omitting any multipliers which do not involve $\theta$ and may therefore be absorbed into the constant of proportionality, the posterior density becomes

$$\begin{aligned}
\pi(\theta|x) &\propto \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\} \\
&\propto \exp\{-\tfrac{1}{2}\theta^2(1/\sigma^2 + 1/\sigma_0^2) + \theta(x/\sigma^2 + \mu_0/\sigma_0^2)\} \\
&= \exp\{-\tfrac{1}{2}\theta^2/\sigma_1^2 + \theta\mu_1/\sigma_1^2\} \\
&\propto \exp\{-\tfrac{1}{2}(\theta-\mu_1)^2/\sigma_1^2\},
\end{aligned} \tag{7}$$

where, in the first and third stages of the argument, terms not involving $\theta$ have been respectively omitted and introduced. The missing constant of proportionality can easily be found from the requirement that $\pi(\theta|x)$ must be a density and therefore integrate to one. It is obviously $(2\pi\sigma_1^2)^{-\frac{1}{2}}$ and so the theorem is proved. (Notice that it is really not necessary to consider the constant at all: it must be such that the integral of $\pi(\theta|x) = 1$, and a constant times (7) is a normal distribution.)

***Corollary.*** *Let* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *be a random sample of size* $n$ *from* $N(\theta, \sigma^2)$, *where* $\sigma^2$ *is known and the prior density of* $\theta$ *is* $N(\mu_0, \sigma_0^2)$. *Then the posterior density of* $\theta$ *is* $N(\mu_n, \sigma_n^2)$, *where*

$$\mu_n = \frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}, \quad \sigma_n^{-2} = n\sigma^{-2} + \sigma_0^{-2}, \tag{8}$$

*and* $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$.

The likelihood is (equation (1))

$$\begin{aligned} p(\mathbf{x}|\theta) &= (2\pi\sigma^2)^{-n/2}\exp\left[-\sum_{i=1}^{n}(x_i-\theta)^2/2\sigma^2\right] \\ &\propto \exp\left[-\tfrac{1}{2}\theta^2(n/\sigma^2) + \theta\bar{x}(n/\sigma^2)\right] \\ &\propto \exp\left[-\tfrac{1}{2}(\bar{x}-\theta)^2\,(n/\sigma^2)\right], \end{aligned} \tag{9}$$

where again terms not involving $\theta$ have been omitted and then introduced. Equation (9) is the same as (5) with $\bar{x}$ for $x$ and $n/\sigma^2$ for $\sigma^2$, apart from a constant. Hence the corollary follows since (8) is the same as (4), again with $\bar{x}$ for $x$ and $n/\sigma^2$ for $\sigma^2$.

*Random sampling*

We have mentioned random samples before (§§1.3, 3.3). They usually arise in one of two situations: either samples are being taken from a large (or infinite) population or repetitions are being made of a measurement of an unknown quantity. In the former situation, if the members of the sample are all drawn according to the rule that each member of the population has the same chance of being in the sample as any other, and the presence of one member in the sample does not affect the chance of any other member being in the sample, then the random variables, $x_i$, corresponding to each sample member will have

a common distribution and be independent, the two conditions for a random sample.† In the second situation the repetitions are made under similar circumstances and one measurement does not influence any other, again ensuring that the two conditions for a random sample are satisfied. The purpose of the repetition in the two cases is the same: to increase one's knowledge, in the first case of the population and in the second case of the unknown quantity—the latter knowledge usually being expressed by saying that the random error of the determination is reduced. In this section we want to see in more detail than previously how the extent of this increase in knowledge can be expressed quantitatively in a special case. To do so it is necessary to express one's knowledge quantitatively; this can be done using probability as a degree of belief (§1.6). Thus our task is to investigate, in a special case, the changes in degrees of belief, due to random sampling. Of course, methods other than random sampling are often used in practice (see, for example, Cochran (1953)) but even with other methods the results for random sampling can be applied with modifications and therefore are basic to any sampling study. Only random sampling will be discussed in this book.

*Likelihood and parameters*

The changes in knowledge take place according to Bayes's theorem, which, in words, says that the posterior probability is proportional to the product of the likelihood and the prior probability. Before considering the theorem and its consequences let us take the three components of the theorem in turn, beginning with the likelihood. The likelihood is equivalently the probability density of the random variables forming the sample and will have the form (1): the product arising from the independence and the multiplication law (equation 3.2.10) and each term involving the same density because of the common distribution. Hence, consideration of the likelihood reduces to consideration of the density of a single member of

† Some writers use the term 'random sample from a population' to mean one taken without replacement (§1.3). In which case our results only apply approximately, though the approximation will be good if the sample is small relative to the population.

the sample. This density is purely a frequency idea, empirically it could be obtained through a histogram (§2.4), but is typically unknown to us. Indeed if it were known then there would be little point in the random sampling: for example, if the measurements were made without bias then the mean value of the distribution would be the quantity being measured, so knowledge of the density implies knowledge of the quantity. But when we say 'unknown', all that is meant is 'not completely known', we almost always know something about it; for example that the density increases steadily with the measurement up to a maximum and then decreases steadily—it is unimodal—or that the density is small outside a limited range—it being very unlikely that the random variable is outside this range. Such knowledge, all part of the 'unknown', consists of degrees of belief about the structure of the density and will be expressed through the prior distribution. It would be of great help if these beliefs could be expressed as a density of a finite number of real variables when the tools developed in the earlier chapters could be used. Otherwise it would be necessary to talk about densities, representing degrees of belief, of functions, namely frequency densities, for which adequate tools are not available. It is therefore usual to suppose that the density of $x$ may be written in the form $f(x|\theta_1, \theta_2, \ldots, \theta_s)$ depending on a number, $s$, of real values $\theta_i$ called parameters; where the function $f$ is known but the parameters are unknown and therefore have to be described by means of a prior distribution. Since we know how to discuss distributions of $s$ real numbers this can be done; for example, by means of their joint density. It is clear that a very wide class of densities can be obtained with a fixed functional form and varying parameters; such a class is called a *family* and later we shall meet a particularly useful class called the *exponential family* (§5.5). In this section we consider only the case of a single parameter, which is restrictive but still important.

Sometimes $f$ is determined by the structure of the problem: for example, suppose that for each member of a random sample from a population we only observe whether an event $A$ has, or has not, happened, and count the number of times it

happens, $x$, say. Then $x$ has a binomial distribution (§2.1) and the only parameter is $\theta = p$, the probability of $A$ on a single trial. Hence the density is known, as binomial, apart from the value of an unknown parameter: the knowledge of the parameter will have to be expressed through a prior distribution. In other situations such reasons do not exist and we have to appeal to other considerations. In the present section the function $f$ is supposed to be the density of a normal distribution with known variance, $\sigma^2$, say, and unknown mean. These are the two parameters of the normal distribution (§2.5). The mean has previously been denoted by $\mu$ but we shall now use $\theta$ to indicate that it is unknown and reserve $\mu$ to denote the true, but unknown, value of the mean. Notice that this true value stays constant throughout the random sampling. The assumption of normality might be reasonable in those cases where past, similar experience has shown that the normal distribution occurs (§3.6). For example, suppose that repeated measurements of a quantity are being made with an instrument of a type which has been in use for many years. Experience with the type might be such that it was known to yield normal distributions and therefore that the same might be true of this particular instrument. If, in addition, the particular instrument had been extensively used in the past, it may have been found to yield results of known, constant accuracy (expressed through the variance or standard deviation). In these circumstances every set of measurements of a single quantity with the instrument could be supposed to have a normal distribution of known variance, only the mean changing with the quantity being measured: if the instrument was free from bias, the mean would be the required value of the quantity. Statistically we say that the scientist is *estimating* the mean of a normal distribution.† This situation could easily occur in routine measurements carried out in the course of inspecting the quality of articles coming off a production line. Often the normal distribution with known variance is assumed with little or no grounds for the normality assumption, simply because it is very easy to handle. That is why it is used here for the first example of quantitative inference.

† Estimation is discussed in §5.2.

*Prior distribution*

The form of the prior distribution will be discussed in more detail in the next section. Here we consider only the meaning of a prior density of $\theta$. We saw, in §1.6, what a prior probability meant: to say that a hypothesis $H$ has prior probability $p$ means that it is considered that a fair bet of $H$ against not-$H$ would be at odds of $p$ to $(1-p)$. We also saw that a density is a function which, when integrated (or summed), gives a probability (§2.2). Hence a prior density means a function which, when integrated, gives the odds at which a fair bet should be made. If $\pi(\theta)$ is a prior density then $\int_0^\infty \pi(\theta)d\theta$ is the prior probability that $\theta$ is positive, and a fair bet that $\theta$ was positive would be at odds of $\int_0^\infty \pi(\theta)d\theta$ to $\int_{-\infty}^0 \pi(\theta)d\theta$ on. In particular, to suppose, as has been done in the statement of the theorem, that $\theta$ has prior density $N(\mu_0, \sigma_0^2)$ means, among other things, that

(i) $\theta$ is believed to be almost certainly within the interval $(\mu_0-3\sigma_0, \mu_0+3\sigma_0)$ and most likely within $(\mu_0-2\sigma_0, \mu_0+2\sigma_0)$ (compare the discussion of the normal distribution in §2.5. We are arbitrarily and conventionally interpreting 'most likely' to mean that the odds against lying outside the interval are 19 to 1).

(ii) $\theta$ is just as likely to be near $\mu_0+\lambda\sigma$ as it is to be near $\mu_0-\lambda\sigma$, for any $\lambda$, and in particular is equally likely to be greater than $\mu_0$ as less than $\mu_0$.

(iii) Within any interval $(\mu_0-\lambda\sigma_0, \mu_0+\lambda\sigma_0)$ the central values are most probable and the further $\theta$ is from the mean, the less likely are values near $\theta$.

*Posterior distribution and precision*

Often these three reasons are held to be sufficient for assuming a normal prior density. But an additional reason is the theorem, which shows that, with a normal likelihood, the posterior distribution is also normal. The extreme simplicity of the result makes it useful in practice, though it should not be used as an excuse for assuming a normal prior distribution when that assumption conflicts with the actual beliefs.

The posterior distribution is, like the prior distribution, one of probability as a degree of belief and because of the normality enables statements like (i)–(iii) above to be made in the light of the datum, the single value of $x$, but with different values of the mean and variance. Let us first consider how these are related to the corresponding values of the prior density and the likelihood; taking the variance first because it is the simpler. We shall call the inverse of the variance, the *precision*. The nomenclature is not standard but is useful and is partly justified by the fact that the larger the variance the greater the spread of the distribution and the larger the intervals in (i) above and therefore the smaller the precision. The second equation in (4) therefore reads:

posterior precision equals the datum precision

plus the prior precision (10)

(this, of course, for normal distributions of datum and prior knowledge and a sample of size 1). The datum precision is the inverse of the random error in the terminology of §3.3. It follows therefore that the posterior precision is necessarily greater than the prior precision and that it can be increased either by an increase in the datum precision (that is by a decrease in the variance of the measurement, or the random error) or by an increase in the prior precision. These statements are all quantitative expressions of rather vaguer ideas that we all possess: their great merit is the numerical form that they assume in the statistician's language. It is part of the statistician's task to measure precision. Notice again that it is the inverse of the variance that occurs naturally here, and not the standard deviation which is used in statements (i)–(iii) above. This agrees with earlier remarks (§§2.4, 3.3) that the variance is easier to work with than the more meaningful standard deviation which can always be obtained by a final square root operation.

The first equation in (4) can also be conveniently written in words provided the idea of a weighted mean is used. A *weighted mean* of two values $a_1$ and $a_2$ with *weights* $w_1$ and $w_2$ is defined as $(w_1a_1+w_2a_2)/(w_1+w_2)$. With equal weights, $w_1 = w_2$, this is the ordinary arithmetic mean. As $w_1$ increases relative to $w_2$ the

weighted mean moves towards $a_1$. Only the ratio of weights is relevant and the definition obviously extends to any number of values. In this terminology

the posterior mean equals the weighted mean of the datum value and the prior mean, weighted with their precisions. (11)

Information about $\theta$ comes from two sources, the datum and the prior knowledge. Equation (11) says how these should be combined. The more precise the datum the greater is the weight attached to it; the more precise the prior knowledge the greater is the weight attached to it. Again this is a quantitative expression of common ideas.

*Small prior precision*

With equations (10) and (11), and the knowledge that the posterior density is normal, revised statements like (i)–(iii) can be made with $\mu_1$ and $\sigma_1$ replacing $\mu_0$ and $\sigma_0$. The most important effect of the datum is that the intervals in these statements will necessarily be narrower, since $\sigma_1 < \sigma_0$; or, expressed differently, the precision will be greater. A most important special case is where the prior precision is very low, or $\sigma_0$ is very large. In the limit as $\sigma_0 \to \infty$ (10) and (11) reduce to saying that the posterior precision and mean are equal to the datum precision and value. Furthermore, both posterior and datum distributions are normal. Consequently there are two results which are quite distinct but which are often confused:

(*a*) the datum, $x$, is normally distributed about a mean $\mu$ with variance $\sigma^2$;

(*b*) the parameter, $\theta$, is normally distributed about a mean $x$ with variance $\sigma^2$.

The first is a statement of frequency probability, the second a statement of (posterior) beliefs. The first is a distribution of $x$, the second a distribution of $\theta$. So they are truly different. But it is very easy to slip from the statement that $x$ lies within three standard deviations of $\mu$ (from (*a*)) to the statement that $\theta$ lies within three standard deviations of $x$ (from (*b*)—cf. (i) above). Scientists (and statisticians) quite often do this and we see that

it is quite all right for them to do so provided the prior precision is low in comparison with the datum precision and they are dealing with normal distributions.

*Precision of random samples*

The corollary establishes similar results for a normal random sample of size $n$ instead of for a single value. It can also usefully be expressed in words by saying:

> a random sample of size $n$ from a normal distribution is equivalent to a single value, equal to the mean of the sample, with $n$ times the precision of a single value. (12)

(An important proviso is that normal distributions are assumed throughout.) The result follows since, as explained in the proof of the corollary, (8) is the same as (4) with $\bar{x}$ for $x$ and $n/\sigma^2$ for $\sigma^{-2}$. The result is related to theorem 3.3.3 that, under the same circumstances, $\mathscr{D}^2(\bar{x}) = \sigma^2/n$, but it goes beyond it because it says that the mean, $\bar{x}$, is equivalent to the whole of the sample. The earlier result merely made a statement about $\bar{x}$, for example that it was a more precise determination than a single observation; the present result says that, with normal distributions it is the most precise determination. This equivalence between $\bar{x}$ and the sample may perhaps be most clearly expressed by considering two scientists both with a random sample of $n$ measurements. Scientist 1 uses the procedure of the corollary. Scientist 2 is careless and only retains the number and mean of his measurements: he then has a single value $\bar{x}$, with mean $\theta$ and variance $\sigma^2/n$ (§3.3), and a normal distribution (§3.5), and can use the theorem. The two scientists end up with the same posterior distribution, provided they had the same prior distribution, so that scientist 2's discarding of the results, except for their number and their mean, has lost him nothing under the assumptions stated. One of a statistician's main tasks used to be called the *reduction of data*, replacing a lot of numbers by a few without losing information, and we see now how this can be done in the special case of a normal distribution of known variance: $n$ values can be replaced by two, $n$ and $\bar{x}$. But remember that this does assume normality, a very important proviso.