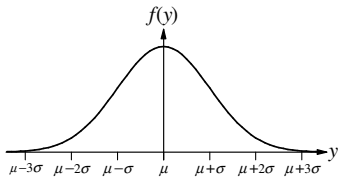
Figure 5.6. MODELLING THE SHAPE OF DATA DISTRIBUTIONS

We are familiar with averages and (data) standard deviations as numerical measures of the *location* and the *variation* of the distribution of a data set; we now turn to one model (or idealization) for the *shape* of a data distribution.

1. Gaussian Distributions

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{y-\mu}{\sigma}\right]^2} \quad ; \quad -\infty < y < \infty \qquad \text{-----(5.6.1)}$$

$$Y \sim G(\mu, \sigma) \qquad \text{-----(5.6.2)}$$



Although a large number of shapes is *possible* for the distribution of a data set, in practice it is observed (by examining suitable histograms, for example) that many distributions fall into one of a limited number of categories. A common shape has a central peak with a roughly symmetrical falling away on either side – for instance, see Figures 3.8 (paper thickness results) and 6.2 (coin weights). A *mathematical model* (or *idealization*) of this shape is the *Gaussian* (*probability*) *distribution*, whose *probability density function* (or *p.d.f.*) is equation (5.6.1) at the right above. The graph of this function is shaped like the cross-section of a bell – in many texts, it is described as a *symmetrical bell-shaped curve*. The equation has two *parameters*, μ and σ, which (independently) determine the position of its centre (μ) and the width of its peak (σ – the *larger* σ, the *wider* the peak). The parameter μ is called the *mean* of the distribution and σ is its (probabilistic) *standard deviation*. The value of μ is the centre of the graph – the y-value of its highest point; the value of σ can be roughly assessed by eye because it is the distance from the mean to either point of inflection. As the diagram indicates, the part of the Gaussian p.d.f. we typically *see* covers an interval of about 3σ either side of μ or about 6σ in total.

Under suitable conditions, probabilities are *estimated* by *proportions*; because proportions are represented by bar *areas* in histograms, and because we model data histograms by probability distributions, it is *area* under a p.d.f. that represents probability.

2. Variates and Random Variables

A *variate* is a characteristic associated with each unit of a population – variates may be *categorical* (like colour or sex or marital status) but our concern here is with variates (like length or weight) that take *numerical* values. In data-based investigating, variate values are usually measured for a *sample* of units selected from an appropriate study population; if the sample contains n units, we denote these (response) variate values by subscripted lower-case Roman letters y_1, y_2, y_3, ..., y_n (or, more compactly, y_j, $j = 1, 2, ... n$).

When we use a probability distribution to model the shape of the distribution of a variate, the variable in the equation of the *model* is called a *random variable*; for example, y in the Gaussian p.d.f. (5.6.1) above is (the value of) a Gaussian random variable. We are familiar with the term 'variable' from algebra and calculus; the addition of the adjective *random* for a variable in a probability distribution can serve to remind us that, by the act of modelling, we have asserted there is a probability associated with each value the variable takes on; these probabilities come, in part, from the *method of selecting* the units which comprise the sample and which yield the data. It is the task of the *Plan* for an investigation (proper selecting *and* proper measuring, etc.) to provide a reasonable basis for regarding variate values as values of a random variable.

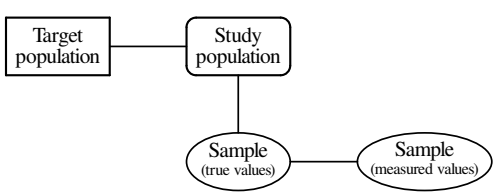
Table 5.6.1

Data	Roman	**Model**	*Italic*
Variate value	y or y_j	Random variable value	y or y_j
		Random variable	Y or Y_j

As summarized in Table 5.6.1 at the right, the usual notation convention is to use a (subscripted) *lower*-case *italic* letter y (or y_j) for the value(s) of a random variable and an *upper*-case *italic* letter Y (or Y_j) for the random variable. When the random variable Y has a *Gaussian* distribution with parameters μ and σ, we write symbolically $Y \sim G(\mu, \sigma)$, shown as equation (5.6.2) above at the right. Letters for random variables are usually chosen from near the *end* of the alphabet, subject to the caveat, to assist problem solving, of using letters whose identity is easy to remember – for instance, W for weight, L for length, T for time; letters near the *beginning* of the alphabet are usually kept for *events*.

The characteristic of a random variable commonly of interest is a *probability*; for instance, the probability the random variable Y is greater than 4 [*i.e.*, symbolically, $\Pr(Y > 4)$], or the probability the random variable Z takes the value z [*i.e.*, $\Pr(Z = z)$].

3. Attributes and Model Parameters: Estimating



The focus of Section 2 above is on *individual* units and their variates, whose values we model as the values of random variables. In this Section, we focus on three *groups* of units:

- the **target population**: the group of units to which the investigator(s) want Answer(s) to the Question(s) to apply;
- the **study population**: a group of units *available* to an investigation;
- the **sample**: the group of units selected from the study population and *actually used* in an investigation.

(*continued overleaf*)

A schema showing these three groups, including the distinction between true and measured variate values, is given at the lower right overleaf (page 5.15); the vertical line in the middle of this schema reminds us the sample is a *sub*set of the study population. [This schema is repeated on the first side (page 5.19) of the following Figure 5.7, and it is then extended and adapted in three more versions on pages 5.21, 5.25, 5.27 and 5.30 – see also the summary on page 5.54.]

Associated with these three *groups* of units are:

∗ **attributes**: quantities defined as a function of the response (and, perhaps, explanatory) variates over the *group* of units.

Familiar (simple) attributes are averages, proportions, medians and standard deviations. The importance of attributes is that:

○ Answer(s) to Questions(s) are usually phrased in terms of attribute values, as introduced in Section 2.2 of the Course Notes;

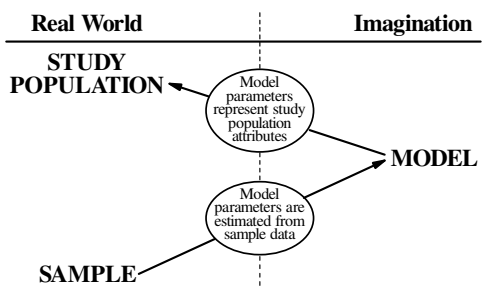
○ *error* is defined in terms of attributes, as discussed in Sections 3.2 to 3.5 of the Course Notes and the following Figure 5.7.

A (probability) *model parameter* is a constant (usually denoted by a *Greek* letter) in a (probability) model that *represents* a study population *attribute*. Model-based methods of analysis in statistics use data from a sample to *estimate* values of model parameters which then represent reasonable values for study population attributes and, hence, for Answer(s) to Question(s) of interest. We distinguish:

∗ a **point** estimate: a *single* value for an estimate; **AND**

∗ an **interval** estimate: an *interval* of values for an estimate.

For a sample of (response) variate values where the Gaussian distribution is an appropriate model, the mean μ is estimated by the sample average $\bar{y}$ and σ is estimated by the sample standard deviation s – these are both *point* estimates. We can think graphically of the process of estimating μ by $\bar{y}$ and σ by s as approximating the histogram of a data set by the Gaussian p.d.f. that has the same 'centre' and the same 'width' as the histogram.



All mathematical models are idealizations and all are products of the intellect and the imagination. It may be helpful to think of the model as a *link* between the *sample* and the *study population*; a pictorial representation of this idea is shown at the right above.

The matters discussed in this Section (and in the Appendix on the last side of the Figure, page 5.18) look ahead to the use of probability models in statistics, which is pursued starting in Chapters 5 and 7 of the Course Notes; our *immediate* concern in this Figure is to become familiar with properties of continuous probability models.

NOTE: 1. To maintain the distinction between the real world (represented by the data) and the model, we use different words – 'average' and 'mean' – for their measures of location; unfortunately, we do not have this option for the two measures of variation, which are both called 'standard deviation'. In the early stages of learning statistics, it is helpful to, at least in our minds, add the respective adjectives 'data' and 'probabilistic' to distinguish the two uses of standard deviation. This terminology is summarized in Table 5.6.2 at the right.

Table 5.6.2:

Attribute	Real World	Model
Location	Average	Mean
Variation	(Data) standard deviation	(Probabilistic) standard deviation

4. Using the Gaussian Distribution

To be able to *use* the Gaussian distribution, we need two skills.

∗ First, we learn to look up a *table* of the standard Gaussian distribution [denoted $G(0,1)$, whose p.d.f. is given as equation (5.6.3) at the right]; such a table furnishes us with *areas* which we use as *probabilities* (see page Ap.1 in Appendix B of the Course Materials).

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad ; \quad -\infty < z < \infty \qquad \text{-----(5.6.3)}$$

$$\frac{Y-\mu}{\sigma} \sim G(0,1) \qquad \text{-----(5.6.4)}$$

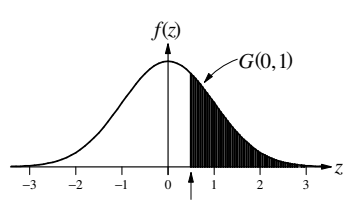
∗ Second, we learn about *standardizing* so we can obtain probabilities for the $G(0,1)$] distribution; if $Y \sim G(\mu, \sigma)$, the result (5.6.4) – in which we *subtract the mean* and *divide by the standard deviation* – at the right above holds; we then can write equation (5.6.5), in which we have standardized to convert a probability for the $G(\mu, \sigma)$ distribution to the equivalent probability for the $G(0,1)$ distribution. The letter Z is commonly used for a random variable with the standard Gaussian distribution.

$$\Pr(Y > y) = \Pr\left(\frac{Y-\mu}{\sigma} > \frac{y-\mu}{\sigma}\right) = \Pr\left[G(0,1) > \frac{y-\mu}{\sigma}\right] \qquad \text{-----(5.6.5)}$$

A numerical illustration of equation (5.6.5) is: if $Y \sim G(8,6)$ and $Z \sim G(0,1)$, then:

$$\Pr(Y > 11) = \Pr\left(\frac{Y-\mu}{\sigma} > \frac{11-8}{6}\right) = \Pr[G(0,1) > 0.5] = \Pr[Z > 0.5];$$

this probability is represented by the area under the standard Gaussian p.d.f. to the *right* of 0.5, shown with solid shading in the diagram at the right. We learned in Figure 5.2 how to look up this probability from the $G(0,1)$ table on page Ap.1 in Appendix B.



(*continued*)

Figure 5.6. MODELLING THE SHAPE OF DATA DISTRIBUTIONS (continued 1)

5. Benefits of Using a Gaussian Model

Three benefits accrue from using the Gaussian distribution.

- First, it provides a *data summary* – instead of having to work with a *data set*, we can convey a number (but usually not *all*) of its essential characteristics by saying merely that the data has a *Gaussian distribution with a specified mean and a specified standard deviation*. This can be a significant convenience, particularly in the case of large data sets. [It is understood, of course, that the Gaussianicity is only *approximate*; also, the location and variation parameters of the Gaussian distribution can take on their values *independently* – that is, the value of one does not influence the value of the other.]
- Second, we can readily *combine* Gaussian distributions; the random variable represented by any *linear combination* of Gaussian random variables has a *Gaussian* distribution, provided the component random variables are *probabilistically independent*. The linear combinations of greatest relevance to statistical methods of data analysis are *sums*, *differences* and *averages* – recall Figure 5.4 on pages 5.11 and 5.12.
- Third, the distribution of a linear combination (such as a sum or an average) of *non*-Gaussian random variables *tends* to a Gaussian distribution. If there is an *infinite* number of components, the distribution of the combination is *exactly* Gaussian, a theoretical result known as the *Central Limit Theorem* – recall the previous Figure 5.5. In *practice*, there are only *finite* combinations, whose distributions therefore exhibit only *approximate* Gaussianicity (unless the components themselves have a Gaussian distribution). How close the approximation is to *exact* Gaussianicity depends on both the number of components in the linear combination and on the shape of their distribution(s) – the more *symmetrical* it is, the smaller the number of components needed for reasonable approximate Gaussianicity.

It is interesting to speculate on the reason(s) behind the mathematical and practical benefits that accrue from Gaussian modelling. It *may* be only coincidence that the mathematics of Gaussian theory works so conveniently in many instances (and *not only* in statistics). Alternatively, $e = 2.71828\ 18284\ldots$ is a quantity that, in a sense, *nature* brings to our attention as, for example, the limit as $n \to \infty$, of $(1+1/n)^n$. It is therefore possible that the wide applicability of Gaussian distribution theory is a reflection of the harmony that results when the mathematics we employ corresponds properly with some aspect(s) of the underlying structure of the physical world. The matter is highlighted in a quotation from Gabriel Lippmann, the French 1908 Nobel Laureate in physics: *Everybody believes in the [Gaussian approximation], the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact*. Paradoxically, *both* beliefs are correct, which reminds us of other seemingly contradictory aspects of experience, like the wave-particle duality of electromagnetic radiation such as light. [The source of the Gabriel Lippmann quotation is the text by Freedman, D., Pisani, R. and R. Purves: *Statistics*. First Edition, W. W. Norton & Company, New York, 1980, page 275.]

6. Using Random Variables in Probability

We *use* random variables to describe (or *model*) quantities that take on different values according to chance; one criterion we use to decide which probability model is appropriate in a particular situation is the degree of agreement between the *shape* of the p.d.f. of the model and the shape of the distribution of the data (as assessed from a histogram, for example). The *stochastic* behaviour of response variates in our models arises because of *equiprobable* (or *random*) *selecting* of units from the relevant population. Some *in*formal descriptions of what is meant by the term 'random variable' (or its value) are:

- ∗ a quantity which takes on different real values according to chance;
- ∗ a numerical outcome of a stochastic phenomenon;
- ∗ a characteristic (like a variate value) which changes from unit to unit in a sample obtained by equiprobable selecting.

In addition to our concern with distribution *shape* in choosing an appropriate probability model, we usually also need to take account of:

- the *mean* of the distribution (or of the random variable, Y say), denoted μ_Y or $E(Y)$;
- the *standard deviation* of the distribution, denoted σ_Y or $s.d.(Y)$.

Formally, a **random variable** is a function which assigns a real number to each point of the sample space (S); *i.e.*, a random variable is a function with domain S and range $\mathbb{R}$ – it is a mapping from the sample space to the real numbers.

7. Distributions Other Than the Gaussian Distribution

Although the Gaussian distribution is widely used as a model for data distributions, other useful models are:

- the *uniform distribution* (which represents the case of equiprobable values, as in equiprobable selecting, and thus has a p.d.f. which is *rectangular*);
- the *exponential distribution* (whose p.d.f. is like an exponential decay, and is used to model failure times);
- the *log Gaussian distribution* (where the *logarithm* of the random variable has a Gaussian distribution; it is used to model biological characteristics which have a lower cut-off at zero but which are not, at least in theory, limited in the *high* values they can take on).

(*continued overleaf*)

In particular methods of data analysis, as well as the Gaussian distribution, we later use of the t, K and χ^2 distributions – see Sections 12.3 and 13.2.2 of the Course Notes and Appendix 3 on pages 13.14 to 13.16 of Figure 13.1 of these Course Materials.

8. Appendix: Response Models

Equation (5.6.2) shows the usual notation in *probability* for stating the distribution of a random variable, Y in this instance. In *statistics*, the *equivalent* expression (5.6.6) at the right, called a *response model*, is more convenient. Its form suggests we think of *two* components making up the random variable Y which, under an appropriate Plan for an investigation [*including* equiprobable selecting (as 'EPS' at the end of the model statement (5.6.6) reminds us)], is used to model the values of a response variate of a unit:

$$Y \sim G(\mu, \sigma) \qquad \text{-----(5.6.2)}$$

$$Y = \mu + R, \quad R \sim G(0, \sigma), \quad \text{EPS} \qquad \text{-----(5.6.6)}$$

- a *structural component* which (in general) models the effect on the response variate of specific explanatory variate(s);
  - in the case of equation (5.6.6), which is the *simplest* response model, the structural component is merely a constant (μ) and contains no explicit explanatory variates;
- a *stochastic component* which models variation about the structural component;
  - in equation (5.6.6), the variation of Y about the structural component of the model (μ in this instance) is modelled by a *Gaussian* distribution with mean 0 and standard deviation σ.

Using the response model (5.6.6) to find an *interval* estimate for each of the model parameters μ and σ, representing the study population *average* response $\bar{\mathbf{Y}}$ and response (data) *standard deviation* $\mathbf{S}$, is pursued in Chapters 12 and 13 of the Course Notes and in Figure 13.1 of these Course Materials.

NOTES: 2. The foregoing discussion uses the word 'model' in two senses:
- the division of Y into a structural component and a stochastic component;
- the Gaussian model for the stochastic component.

Only the *second* of these models involves probability.

3. The use of a *probability* distribution (here, the *Gaussian* distribution) to model the stochastic component of a response model illustrates one reason why probability is useful in statistics.

4. Equation (5.6.6) can be extended, as shown in equation (5.6.7) at the right, to model response variate values from a sample of n units selected equiprobably from a study population; like the requirement for equiprobable selecting, the modelling assumption of *probabilistic independence* of the Y_js has implications for Plan components which address how the data are to be collected.

$$Y_j = \mu + R_j, \quad j = 1, 2, \ldots, \mathrm{n}, \quad R_j \sim G(0, \sigma), \quad \text{independent}, \quad \text{EPS} \qquad \text{-----(5.6.7)}$$

5. The *formal* definitions of the symbols in equation (5.6.7) are:

   Y_j is a random variable whose distribution represents the possible values of the measured response variate ($\mathbf{Y}$) for the jth unit in the sample of n units selected equiprobably from the study population of $\mathbf{N}$ units (usually $\mathrm{n} \ll \mathbf{N}$), if the selecting and measuring processes were to be repeated over and over.

   μ is a model parameter (called the *mean*) which represents the *average* ($\bar{\mathbf{Y}}$) of the measured response variate ($\mathbf{Y}$) of the units of the study population.

   R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate for the jth unit in the sample of n units selected equiprobably from the study population, if the selecting and measuring processes were to be repeated over and over.

   σ the (probabilistic) *standard deviation* of the Gaussian model for the distribution of the residual, is a model parameter which represents the (data) *standard deviation* ($\mathbf{S}$) of the measured response variate ($\mathbf{Y}$) of the units of the study population; this (data) standard deviation (and, hence, σ) *quantifies* the *variation* of the measured response variate of the units of the study population – as this variation increases, so does the study population (data) standard deviation (and, hence, so does σ).

   In these definitions (and elsewhere in this Figure 5.6), 'study population' should actually be 'respondent population', as discussed starting in Section 5 on page 5.24 in the following Figure 5.7 – see also Figure 5.9 on pages 5.97 and 5.98. [This refinement in terminology is *absent* from the STAT 231 Course Notes but, for consistency in these Course Materials, *is* used after Figure 5.7 (*e.g.*, in Figures 5.8 and 13.1).

6. Another way of writing the response model (5.6.6) is equation (5.6.8), where the model for R is now *standard* Gaussian.

$$Y = \mu + \sigma R, \quad R \sim G(0, 1), \quad \text{EPS} \qquad \text{-----(5.6.8)}$$

[1] The late Dr. George E.P. Box, a respected U.S. statistician, coined the maxim: *All models are wrong, some are useful.* Give an explanation, which could be understood by an intelligent but non-technical audience, of what Dr. Box meant.