Figure 5.5. PROBABILITY MODELLING: The Central Limit Theorem Approximation

Figures 5.1 to 5.4 introduce Gaussian distributions and random variables, probabilistic matters useful for statistical modelling.

* In Figure 5.4, a random variable which is a linear combination (like a sum, a difference or an average) of Gaussian (probabilistically independent) random variables also has a Gaussian distribution and its mean and standard deviation can be expressed in terms of the mean(s) and standard deviation(s) of the random variables that make up the linear combination. For example, for n probabilistically independent $G(\mu, \sigma)$ random variables, as equations (5.5.1) and (5.5.2) below remind us:
  - their sum (T) has a *Gaussian* distribution with mean $n\mu$ and standard deviation $\sqrt{n}\sigma$; $T \sim G(n\mu, \sqrt{n}\sigma)$ -----(5.5.1)
  - their *average* has a *Gaussian* distribution with the same mean μ as the individual random variables but a standard deviation that is smaller by a factor of $\sqrt{1/n}$. $\overline{Y} \sim G(\mu, \sigma\sqrt{\tfrac{1}{n}})$ -----(5.5.2)

In this Figure 5.5, we discuss relaxing the requirement for *Gaussianicity* of the distribution of the individual random variables in a linear combination like a sum or an average; previous results for means and standard deviations from Figure 5.4 carry over *un*changed.

* In the following Figure 5.6, we develop the idea of using the Gaussian distribution to model the *shape* of appropriate data distributions; this idea can be summarized by the probability statement (5.5.3) above. $Y \sim G(\mu, \sigma)$ -----(5.5.3)

1. The Central Limit Theorem (abbreviated CLT)

We state without proof this famous result of probability theory as: if the random variables $Y_1, Y_2, Y_3, \ldots.., Y_n$ each have mean μ and standard deviation σ, and if the random variable $T = Y_1 + Y_2 + Y_3 + \ldots.. + Y_n$, then:

- the standardized form of T, $(T - n\mu)/(\sqrt{n}\sigma)$, has a *standard Gaussian p.d.f. in the limit as* $n \to \infty$,
- the standardized form of $\overline{Y} \equiv T/n$, $(\overline{Y} - \mu)/(\sigma/\sqrt{n})$, has a *standard Gaussian p.d.f. in the limit as* $n \to \infty$.

The CLT provides an *exact* result when n is *infinite*; we use it as the justification of an *approximate* result when n is *finite*: a sum (T) or average ($\overline{Y}$) of n random variables Y_j has approximately a Gaussian distribution *regardless of the distribution of the* Y_js; these two results are stated symbolically at the right as equations (5.5.4) and (5.5.5).

$T \div G(n\mu, \sqrt{n}\sigma)$ -----(5.5.4)

$\overline{Y} \div G(\mu, \sigma\sqrt{\tfrac{1}{n}})$ -----(5.5.5)

The *accuracy* of this approximate Gaussianicity of a sum or average depends on two factors:

- the value of n – the *larger* the value, the *better* the approximation;
- the shape of the distribution(s) of the Y_js – the *more symmetrical* they are, the *better* the approximation.

Although it is beyond this Figure to discuss how to assess quantitatively the accuracy of the Gaussian approximation from the CLT in a *particular* situation, general guidelines are:

- if the Y_js have a symmetrical distribution(s), an adequate approximation for many practical purposes (*e.g.*, probabilities accurate to a few percent or better) can be obtained with n as small as 20 to 50;
- with highly asymmetric distribution(s) for the Y_js, the approximation may be of poor accuracy with n as large as 50,000.

Two other restrictions on the use of the Gaussian approximation from the CLT are:

○ the standard deviation(s) of the Y_js must be finite; [this is more of theoretical interest than of practical concern]
○ the Y_js need not be *strictly* probabilistically independent, but they must not be *too* strongly associated.

We use the CLT approximation to estimate probabilities, usually in the context of questions of *statistical* interest, but the idea that a response variate is observed to have approximately a Gaussian distribution is of more *general* interest; for example, the fact that human *heights*, for instance, are quite closely modelled by a Gaussian distribution has been taken as evidence that *many*, not just a few, explanatory variates determine a person's adult height.

NOTE: 1. Figure 5.4 deals with *three* linear combinations (sums, differences and averages) of random variables, but only sums and averages are discussed above. The reason is *differences* usually only involve *two* random variables, so the value of n is too small for the CLT to provide reasonable approximate Gaussianicity of the random variable involving a difference, except when the *individual* random variables have a Gaussian distribution (as they do in Figure 5.4).

Example 5.5.1: (a) An insurance company calculates premiums to many decimal places and then rounds them to the nearest dollar. By modelling the fractional parts of 30,000 premiums by a continuous uniform distribution on (–½, ½], find the approximate probability the rounding alters the *total* amount of these premiums by more than \$50; by more than \$100. [*0.3174 ≃ 0.32*; *0.0456 ≃ 0.046*]

(b) Suppose the premiums in (a) are *first* rounded to the nearest cent and *then* rounded to the nearest dollar, with 50¢ being rounded upwards. Find approximate values for the same probabilities as in (a). [*0.97728 ≃ 0.98*; *0.8413 ≃ 0.84*]

(*continued overleaf*)

Solution: (a) Let the random variable Y_j represent the amount (in dollars) by which the jth premium changes; for example, if the first premium is calculated as \$206.8176 and rounded to \$207, $Y_1 = +0.1824$ dollars.

We use the model: $Y_j \sim U(-½, ½]$ for which: $f(y_j) = 1 \quad ; \quad 0.5 < y_j \le 0.5$;

from Figure 5.11 of the STAT 220 Course Materials, we know that: $E(Y_j) = 0, \quad s.d.(Y_j) = \frac{1}{2\sqrt{3}} \simeq 0.288\,675$ dollars.

The change due to rounding in the *total* premium amount is given by the random variable: $T = \sum_{j=1}^{30,000} Y_j$;

then: $E(T) = \sum_{j=1}^{30,000} E(Y_j) = 0, \quad s.d.(T) = \sqrt{\sum_{j=1}^{30,000} [s.d.(Y_j)]^2} = \sqrt{30,000(\frac{1}{2\sqrt{3}})^2} = 50$ dollars.

Hence, using the CLT approximation: $T \div G(0, 50)$,

so that: $\Pr(|T| > 50) \simeq 2 \times \Pr[G(0,1) > 1] = 2 \times 0.1587 = 0.3174 \simeq 0.32.$

With n as *large* as 30,000 and the Y_js having a *symmetrical* distribution, we expect *good* accuracy for the approximate Gaussianicity of T from the CLT; the final answer should therefore be close to the true probability.

(b) To understand the effect of rounding in *two* stages, we examine particular premium values, as in Table 5.5.1 at the right, where the – and + signs in the second ('Premium') column mean 'infinitesimally below' and 'infinitesimally above'.

Table 5.5.1

	Premium	**Rounding**	Y_j
(a)	\$265.50–	Down	−0.5
	\$265.50+	Up	+0.5
(b)	\$265.495–	Down	−0.495
	\$265.495+	Up	+0.505

We see that, in (a), the break points are ±0.5 dollars so we use a $U(-½, ½]$ model; in (b), the break points are −0.495 and +0.505 dollars so we must use a $U(-0.495, 0.505]$ model.

We now have, from Figure 5.11: $E(Y_j) = 0.005, \quad s.d.(Y_j) = \frac{1}{2\sqrt{3}} \simeq 0.288\,675$ dollars.

Based on the solution above for (a), the solution for (b) [and the second probability in (a)] should be completed as exercises.

NOTES: 2. A noteworthy feature of Example 5.5.1 is the substantial difference in the probabilities in (a) and (b) resulting from what might appear to be an inconsequential change in the rounding procedure.

3. The problem statement in (b) specifes 50¢ is to be rounded *upwards*; explain briefly how this affects the solution.

Example 5.5.2: The lifetimes of certain electronic components are independent and can be modelled by an exponential distribution with a mean of 1,000 hours. Use the Central Limit Theorem to find the approximate probability the *average* lifetime of ten of the components, chosen at random, exceeds 1,500 hours. Explain briefly why the CLT approximation is expected to be of poor accuracy in this instance. [*0.05692 ≃ 0.06*]

Solution: Let the random variable T_j represent the lifetime (in hours) of the jth component.

We use the model: $T_j \sim Exp(\theta = 1{,}000)$;

from Figure 5.12 of the STAT 220 Course Materials, we know that: $E(T_j) = s.d.(T_j) = 1{,}000$ hours.

The *average* lifetime of 10 components is given by the random variable: $\overline{T} = \frac{\sum_{j=1}^{10} T_j}{10}$;

then: $E(\overline{T}) = E(T_j) = 1{,}000; \quad s.d.(\overline{T}) = s.d.(T_j)\sqrt{\frac{1}{10}} = 1{,}000\sqrt{\frac{1}{10}} = 100\sqrt{10} \simeq 316.2$ hours.

Hence, using the CLT approximation: $\overline{T} \div G(1{,}000, 100\sqrt{10})$,

so that: $\Pr(\overline{T} > 1{,}500) \simeq \Pr[G(0,1) > \sqrt{2.5}] = 0.05692 \simeq 0.06.$

With n as *small* as 10 and the T_js having a *very* asymmetric distribution, we expect *poor* accuracy for the approximate Gaussianicity of $\overline{T}$ from the CLT; the final answer may therefore *not* be close to the true probability.

NOTES: 4. Examples 5.5.1 and 5.5.2 involve *continuous* distributions – the continuous uniform and exponential – for the Y_js and T_js; however, the CLT approximation is also applicable when Y_j and T_j are *discrete* random variables.

5. In equations (5.5.1) and (5.5.4) overleaf on page 5.13, it is natural in this Figure to write the standard deviation of T as $\sqrt{n}\sigma$, but it is also useful to think of it as $\sigma\sqrt{n}$, just as we write the standard deviation of $\overline{Y}$ as $\sigma\sqrt{1/n}$. Writing standard deviations this way prepares us for expressing the standard deviation of an estimator in statistical inference as the model standard deviation (usually denoted σ) multiplied by an expression involving a square root.