Figure 6.2. PROBABILITY MODELLING: The Sample Average and Control Charts

Program 18 in: *Against All Odds: Inside Statistics*

Sample counts and proportions, which are discussed in Program 17, are important sample statistics. This program presents the facts about another very important statistic, the sample average $\bar{y}$. Suppose we have a sample of size n obtained by equiprobable (or simple random) selecting [abbreviated EPS (or SRS)] from a large population with average (or model mean) μ and standard deviation σ. Some essential properties of the sampling distribution of the random variable $\bar{Y}$, repesenting the average $\bar{y}$ of the sample obtained by EPS (or SRS), are:

- **Mean:** $\mu(\bar{Y}) \equiv \mu_{\bar{Y}} \equiv E(\bar{Y}) = \mu$;
- **S.d.** and **Variance:** $\sigma(\bar{Y}) \equiv \sigma_{\bar{Y}} \equiv s.d.(\bar{Y}) = \sigma\sqrt{\frac{1}{n}}$ and $\sigma^2(\bar{Y}) \equiv \sigma^2_{\bar{Y}} \equiv var(\bar{Y}) = [\sigma\sqrt{\frac{1}{n}}]^2 = \frac{\sigma^2}{n}$;
- **Distribution:** If the population distribution is normal, $\bar{Y}$ has a *normal* distribution; *i.e.*, $\bar{Y} \sim N(\mu, \sigma\sqrt{\frac{1}{n}})$
- **Distribution:** *Regardless* of the population distribution, the distribution of $\bar{Y}$ is *approximately* normal when the sample size n is large, as a consequence of the *Central Limit Theorem*. *i.e.*, $\bar{Y} \div N(\mu, \sigma\sqrt{\frac{1}{n}})$

The first property says that the sample average $\bar{Y}$ is an *unbiased estimator* of the population average (or model mean) μ; the second property says that the variation of a sample average (or model mean) *de*creases as the sample size *in*creases. The third and fourth properties both concern the *form* of the distribution of $\bar{Y}$. It turns out that *any* sum, difference or average of (probabilistically) *independent* normal random variables also has a normal distribution. This leads to the third fact: if the population has a normal distribution $N(\mu, \sigma)$, then $\bar{Y}$ has a $N(\mu, \sigma/\sqrt{n})$ distribution. What if the population does *not* have a normal distribution? The consequence of the Central Limit Theorem is that, for large n, the sampling distribution of $\bar{Y}$ is *approximately* $N(\mu, \sigma/\sqrt{n})$ for any population with finite standard deviation σ; this is true *even* if the population distribution is *discrete*.

In the video, the Central Limit Theorem is applied to playing roulette in a casino. Because the standard deviation σ of the payoff on a *single* play is large, the standard deviation $\sigma/\sqrt{n}$ is still large for $n = 50$ or so bets. Gambling is exciting for the gambler because the outcome of an evening's betting is uncertain. By contrast, the casino bets a *very* large number of times; hence, $\sigma/\sqrt{n}$ for the casino is small and the average payoff $\bar{y}$ will be close to the population average (or model mean) μ. This average favours the casino a little so, in the long run, it makes money at a *predictable* rate.

These facts about the distribution of $\bar{Y}$ are applied in *statistical process control* to help monitor and improve the quality of manufactured goods. The idea is to watch the manufacturing process to catch changes *early* rather than waiting to inspect the product at the *end*. A process that can be measured over time is said to be *in control* if its pattern of variation (its probability distribution) is the *same at all times*. A process that is in control is operating under stable conditions; *i.e.*, it is said to be a *stable process*.

A $\bar{y}$ (or $\bar{x}$) *control chart* (also called a *Shewhart chart* or a *control chart for averages*) is a graph of sample averages plotted against the time order of the samples, to show whether the level of the process is changing over time. The chart has a solid *centre line* at the target value μ of the process average and dashed *control limits* at $\mu \pm 3\sigma/\sqrt{n}$. The control limits include the range of variation in $\bar{y}$ we expect to see in a normally operating process. A value outside these limits suggests that the process has been disturbed by some additional source of variation that should be located and fixed. In this way, a $\bar{y}$ chart helps us decide if a process is in control with average μ and standard deviation σ. In practice, an out-of-control signal results when:

- any one point falls outside the control limits; *i.e.*, beyond 3 standard deviations on *either* side of the centre line; **OR**
- two of three consecutive points lie between 2 an 3 standard deviations on the *same* side of the centre line; **OR**
- four of five consecutive points lie between 1 and 2 standard deviations on the *same* side of the centre line; **OR**
- there is a *run* of 8 (or, sometimes, 9) consecutive points on the *same* side of the centre line.

In the video, you will see control charts being used to monitor the salt content of potato chips.

1. The symbols $\bar{y}$ and $\bar{Y}$ (or $\bar{x}$ and $\bar{X}$) are used in a number of places in the summary of the video contents given above. Explain carefully the (statistical) *distinction(s)* between the lower and upper case symbols and their use(s).

2. The first sentence of the second paragraph above refers to an *unbiased estimator*; explain the meaning of this term in your own words.
   - From what sources does *inaccuracy* arise? Explain briefly.
     - Which is usually of greater concern in practice: estimating bias or the sources of inaccuracy you identified? Explain briefly.

(*continued overleaf*)

3 In your own words, explain what the *Central Limit Theorem* says; make clear whether it is an *exact* or an *approximate* result.

4 The first sentence of the third paragraph of the summary of the video contents given overleaf on page 6.13 states that .... *the Central Limit Theorem is applied to playing roulette in a casino.* Explain the precise sense in which the Theorem is being applied.

5 Explain briefly to what the *approximation* refers in the *lower* of the two symbolic expressions for $\overline{Y}$ given overleaf on page 6.13 at the upper right; distinguish clearly between what is, and is not, approximate, and indicate the assumption(s) involved.

6 The second-last paragraph of the summary of the video contents given overleaf on page 6.13 states ... *the idea is .... to catch changes* early *rather than waiting to inspect the product at the* end. What are the *advantage(s)* of the former approach (charting) over the latter (inpection) as a method of quality assurance?

7 The second-last paragraph of the summary overleaf on page 6.13 defines a process that is *in control*; show this definition in *pictorial* (or *graphical*) form.

- Construct a corresponding pictorial display for a process that is *out of control*.

8 What does n represent in the control limits of $\mu \pm 3\sigma\sqrt{\frac{1}{n}}$, given in the third line of the last paragraph of the summary overleaf on page 6.13?

9 For a stable process, find the probability of an observation (*i.e.*, a sample average) lying *outside* the control limits of $\mu \pm 3\sigma/\sqrt{n}$. Show your method of calculation and state any assumption(s) involved.

- Repeat the probability calculation for a run of 8 and of 9 consecutive points on the *same* side of the centre line; again show your method and assumption(s).
- Repeat the probability calculation for each of the *other* two out-of-control signals described near the bottom of the summary of the video contents given overleaf on page 6.13.
  - On the basis of your calculations, arrange the four out-of-control signals in order from most to least sensitive.