Chapter 6

Central limit theorems

6.1 Overview

Recall that a random variable Z is said to have a ***standard normal*** distribution, denoted by $N(0,1)$, if it has a continuous distribution with density

$$\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2) \qquad \text{for } -\infty < z < \infty.$$

That is, for all intervals $[a,b]$,

$$\mathbb{P}\{a \le Z \le b\} = \int_a^b \phi(z)\, dz,$$

and, for each subset A of the real line, $\mathbb{P}\{Z \in A\} = \int_A \phi(z)\, dz$. In particular, for each fixed b we must have $\mathbb{P}\{Z = b\} = \int_b^b \phi(z)\, dz = 0$.

More generally, for $\mu \in \mathbb{R}$ and $\sigma > 0$, a random variable X is said to have a $N(\mu, \sigma^2)$ ***distribution*** if $(X-\mu)/\sigma$ has a $N(0,1)$ distribution. That is,

$$\begin{aligned}\mathbb{P}\{a \le X \le b\} &= \mathbb{P}\{(a-\mu)/\sigma \le (X-\mu)/\sigma) \le (b-\mu)/\sigma\} \\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \phi(z)\, dz \\ &= \int_a^b f_{\mu,\sigma}(x)\, dx\end{aligned}$$

where

$$f_{\mu,\sigma}(x) := \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{for } -\infty < x < \infty.$$

In other words, X has a continuous distribution with density $f_{\mu,\sigma}(x)$.

Remark. In Chapter 7 you will see that if Z has a $N(0,1)$ distribution then $\mathbb{E}Z = 0$ and $\text{var}(Z) = 1$. Consequently, if X has a $N(\mu, \sigma^2)$ distribution then $\mathbb{E}X = \mu$ and $\text{var}(X) = \sigma^2$.

The normal approximation to the Binomial distribution also implies a normal approximation for the distribution of some other random variables.

Example <6.1> A normal approximation for a sample median

The normal approximation to the Binomial is just one example of a general phenomenon corresponding to the mathematical result known as the ***central limit theorem***. Roughly stated, the theorem asserts:

If X can be written as a sum of a large number of relatively small, independent random variables, and if $\mathbb{E}X = \mu$ and $\text{var}(X) = \sigma^2$, then the standardized variable $(X - \mu)/\sigma$ has approximately a standard normal distribution. Equivalently, X is approximately $N(\mu, \sigma^2)$ distributed.

If you are interested in the reasons behind the success of normal approximation, see the Appendix to Chapter 8 for an outline of a proof of the central limit theorem.

The normal distribution has many agreeable properties that make it easy to work with. Many statistical procedures have been developed under normality assumptions, with occasional offhand references to the central limit theorem to mollify anyone who doubts that all distributions are normal. That said, let me also note that modern theory has been much concerned with possible harmful effects of unwarranted assumptions such as normality. The modern fix often substitutes huge amounts of computing for neat, closed-form, analytic expressions; but normality still lurks behind some of the modern data analytic tools.

Example <6.2> A hidden normal approximation—the boxplot

The normal approximation is heavily used to give an estimate of variability for the results from sampling.

Example <6.3> Normal approximations for sample means

6.2 The examples

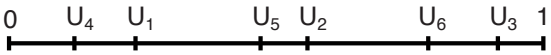
<6.1> **Example.** Suppose $U_1, \ldots, U_n$ are independent random variables each distributed Uniform$(0,1)$. That is,

$$\mathbb{P}\{a \le U_i \le b\} = b - a \qquad \text{for all } 0 < a \le b < 1.$$

The corresponding density function is $f(z) = \mathbf{1}\{0 < z < 1\}$.

For simplicity suppose n is even, $n = 2k$. The sample median M_n is defined as the kth smallest when the U_i's are arranged in increasing order.

Remark. Some authors would define M_n as the $(k+1)$st smallest or as some value between the kth and $(k+1)$st. It doesn't make much difference when n is large.



For example, if $n = 6$ and the U_i's are as shown then M_n would be equal to U_5. For another realization it would probably be equal to another U_i.

Now consider any fixed y in $(0,1)$. Write N_y for the number of U_i's that are $\le y$. More formally,

$$N_y = \sum\nolimits_{i \le n} \mathbf{1}\{U_i \le y\}.$$

The random variable N_y counts the number of "successes" (the number of U_i's that are $\le y$) in n independent trials; N_y has Bin(n, y) distribution, with expected value ny and variance $ny(1-y)$. The key thing to notice is:

$$N_y \ge k \qquad \text{iff} \quad \text{"at least } k \text{ of the } U_i\text{'s are} \le y \quad \text{iff} \quad M_n \le y.$$

Thus

$$\begin{aligned}\mathbb{P}\{M_n \le y\} &= \mathbb{P}\{N_y \ge k\} \\ &= \mathbb{P}\left\{ \frac{N_y - ny}{\sqrt{ny(1-y)}} \ge \frac{k - ny}{\sqrt{ny(1-y)}} \right\}.\end{aligned}$$

Use the normal approximation for the distribution of the standardized variable $(N_y - ny)/\sqrt{ny(1-y)}$ to deduce that the last probability is approximately equal to

$$\int_\gamma^\infty \phi(y)\, dy = 1 - \Phi(\gamma) \qquad \text{where } \gamma := (k - ny)/\sqrt{ny(1-y)}.$$

Now consider a special value, $y = (1 + x/\sqrt{n})/2$, for a fixed x. When n is large enough we certainly have $y \in (0, 1)$. This choice also gives

$$ny(1 - y) = \frac{n}{4}\left(1 - \frac{x^2}{n}\right) \approx \frac{n}{4}$$

and

$$k - ny = -x\sqrt{n}/2,$$

implying $\gamma \approx -x$ and

$$\mathbb{P}\{M_n \le (1 + x/\sqrt{n})/2\} \approx 1 - \Phi(-x) = \Phi(x).$$

For the last equality I have used the symmetry of ϕ around zero to deduce that $\int_{-x}^{\infty} \phi(y)\, dy = \int_{-\infty}^{x} \phi(y)\, dy$.

Put another way,

$$\mathbb{P}\{2\sqrt{n}(M_n - 1/2) \le x\} \approx \Phi(x)$$

which shows that $2\sqrt{n}(M_n - 1/2)$ is approximately $N(0, 1)$ distributed.

Remark. It might be more convincing to use the approximation twice, first with $x = b$ and then with $x = a$, where $a < b$, then subtract.

That is, M_n has approximately a $N(1/2, 1/(4n))$ distribution. □

<6.2> **Example.** The boxplot provides a convenient way of summarizing data (such as grades in Statistics 241/541). The method is:

(i) arrange the data in increasing order

(ii) find the split points

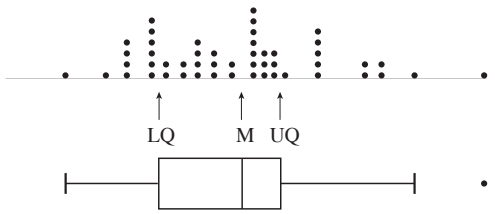
LQ = lower quartile: 25% of the data smaller than LQ
M = median: 50% of the data smaller than M
UQ = upper quartile: 75% of the data smaller than UQ

(iii) calculate IQR (= inter-quartile range) = UQ−LQ

(iv) draw a box with ends at LQ and UQ, and a dot or a line at M

(v) draw whiskers out to UQ + (1.5 × IQR) and LQ − (1.5 × IQR), but then trim them back to the most extreme data point in those ranges

(vi) draw dots for each individual data point outside the box and whiskers (There are various ways to deal with cases where the number of observations is not a multiple of four, or where there are ties, or ...)



Where does the $1.5 \times IQR$ come from? Consider n independent observations from a $N(\mu, \sigma^2)$ distribution. The proportion of observations smaller than any fixed x should be approximately equal to $\mathbb{P}\{W \le x\}$, where W has a $N(\mu, \sigma^2)$ distribution. From normal tables (or a computer),

$$\mathbb{P}\{W \le \mu + .675\sigma\} \approx .75 \qquad \text{and} \qquad \mathbb{P}\{W \le \mu - .675\sigma\} \approx .25$$

and, of course, $\mathbb{P}\{W \le \mu\} = .5$. For the sample we should expect

$$\text{LQ} \approx \mu - .675\sigma \qquad \text{and} \qquad \text{UQ} \approx \mu + .675\sigma \qquad \text{and} \qquad \text{M} \approx \mu$$

and consequently, IQR $\approx 1.35\sigma$. Check that $0.675 + (1.5 \times 1.35) = 2.70$. Before trimming, the whiskers should approximately reach to the ends of the range $\mu \pm 2.70\sigma$. From computer (or tables),

$$\mathbb{P}\{W \le \mu - 2.70\sigma\} = \mathbb{P}\{W \ge \mu + 2.70\sigma\} = .003$$

Only about 0.6% of the sample should be out beyond the whiskers. □

<6.3> **Example.** Chapter 4 gave the expected value and variance of a sample mean $\bar{Y}$ for a sample of size n (with replacement) from a finite population labelled $1, \ldots, N$ with "values of interest" $y_1, y_2, \ldots, y_N$:

$$\mathbb{E}\overline{Y} = \overline{y} = \frac{1}{N}\sum\nolimits_{i=1}^{N} y_i.$$

For sampling with replacement,

$$\text{var}(\overline{Y}) = \sigma^2/n \qquad \text{where } \sigma^2 = \textstyle\sum_{i=1}^{N} (y_i - \overline{y})^2 / N.$$

The standardized random variable $(\overline{Y} - \overline{y})/\sqrt{\sigma^2/n}$ is well approximated by the $N(0,1)$. Thus

$$\mathbb{P}\left\{-\frac{1.96\sigma}{\sqrt{n}} \le \overline{Y} - \overline{y} \le \frac{1.96\sigma}{\sqrt{n}}\right\} \approx \Phi(1.96) - \Phi(-1.96) \approx 0.95.$$

Before we sample, we can assert that we have about a 95% chance of getting a value of $\overline{Y}$ in the range $\overline{y} \pm 1.96\sigma/\sqrt{n}$. (For the post-sampling interpretation of the approximation, you should take Statistics 242/542.)

Of course, we would not know the value σ, so it must be estimated. How?

For sampling without replacement, the variance of the sample mean is multiplied by the correction factor $(N-n)/(N-1)$. The sample mean is no longer an average of many *independent* summands, but the normal approximation can still be used. (The explanation would take me too far beyond 241/541.) □