UNIVERSITY OF LONDON

Probability and Statistics: To p, or not to p?

Module Leader: Dr James Abdey

5.5 The central limit theorem

We have discussed (in Section 4.5) the very convenient result that if a random sample comes from a normally-distributed population, the sampling distribution of $\bar{X}$ is also normal. How about sampling distributions of $\bar{X}$ from other populations?

For this, we can use a remarkable mathematical result, the **central limit theorem (CLT)**. In essence, the CLT states that the normal sampling distribution of $\bar{X}$ which holds *exactly* for random samples from a normal distribution, also holds *approximately* for random samples from *nearly any* distribution.

The CLT applies to 'nearly any' distribution because it requires that the variance of the population distribution is finite. If it is not, the CLT does not hold. However, such distributions are not common.

Suppose that $\{X_1, X_2, \ldots, X_n\}$ is a random sample from a population distribution which has mean $\mathrm{E}(X_i) = \mu < \infty$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$, that is with a finite mean and finite variance. Let $\bar{X}_n$ denote the sample mean calculated from a random sample of size n, then:

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z\right) = \Phi(z)$$

for any z, where $\Phi(z) = P(Z \leq z)$ denotes a cumulative probability of the standard normal distribution.

The '$\lim_{n\to\infty}$' indicates that this is an **asymptotic** result, i.e. one which holds increasingly well as n increases, and exactly when the sample size is infinite.

In less formal language, the CLT says that for a random sample from *nearly any* distribution with mean μ and variance σ^2 then:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

approximately, when n is sufficiently large. We can then say that $\bar{X}$ is **asymptotically normally distributed** with mean μ and variance σ^2/n.

The wide reach of the CLT

It may appear that the CLT is still somewhat limited, in that it applies only to sample means calculated from random samples. However, this is not really true, for two main reasons.

- There are more general versions of the CLT which do not require the observations X_i to be independent and identically distributed (IID).
- Even the basic version applies very widely, when we realise that the 'X' can also be a function of the original variables in the data. For example, if X and Y are random variables in the sample, we can also apply the CLT to:

$$\sum_{i=1}^{n} \frac{\log(X_i)}{n} \quad \text{or} \quad \sum_{i=1}^{n} \frac{X_i Y_i}{n}.$$

Therefore, the CLT can also be used to derive sampling distributions for many statistics which do not initially look at all like $\bar{X}$ for a single random variable in a random sample. You may get to do this in future courses.

How large is 'large n'?

The larger the sample size n, the better the normal approximation provided by the CLT is. In practice, we have various rules-of-thumb for what is 'large enough' for the approximation to be 'accurate enough'. This also depends on the population distribution of X_i. For example:

- for symmetric distributions, even small n is enough
- for very skewed distributions, larger n is required.

For many distributions, $n > 50$ is sufficient for the approximation to be reasonably accurate.

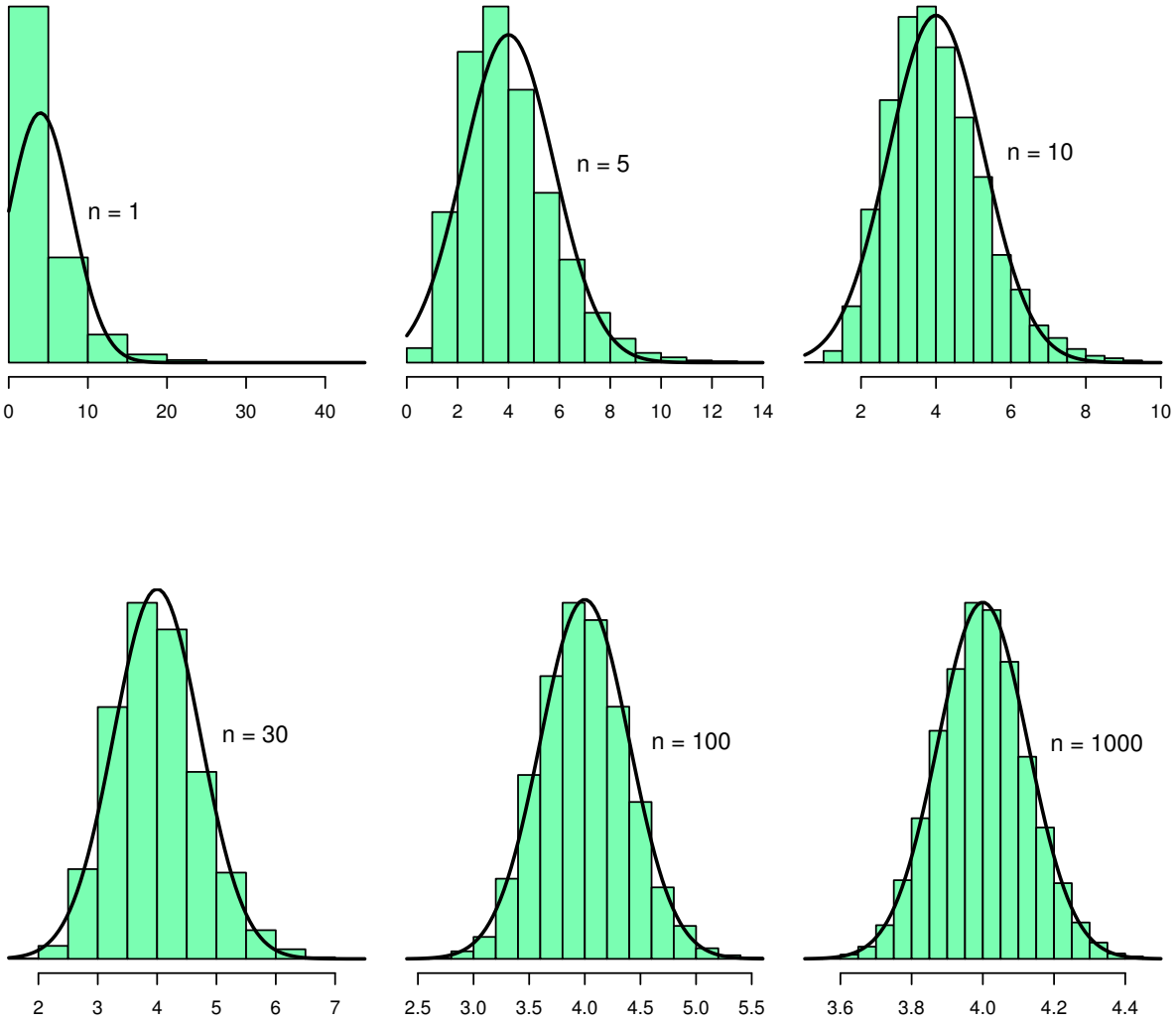
Example

In the first case, we simulate random samples of sizes:

$$n = 1, 5, 10, 30, 100 \text{ and } 1000$$

from the Exponential(0.25) distribution (for which $\mu = 4$ and $\sigma^2 = 16$). This is clearly a skewed distribution, as shown by the histogram for $n = 1$ in the figure below.

10,000 independent random samples of each size were generated. Histograms of the values of $\bar{X}$ in these random samples are shown. Each plot also shows the approximating normal distribution, $N(4, 16/n)$. The normal approximation is reasonably good already for $n = 30$, very good for $n = 100$, and practically perfect for $n = 1000$.



Example

In the second case, we simulate 10,000 independent random samples of sizes:

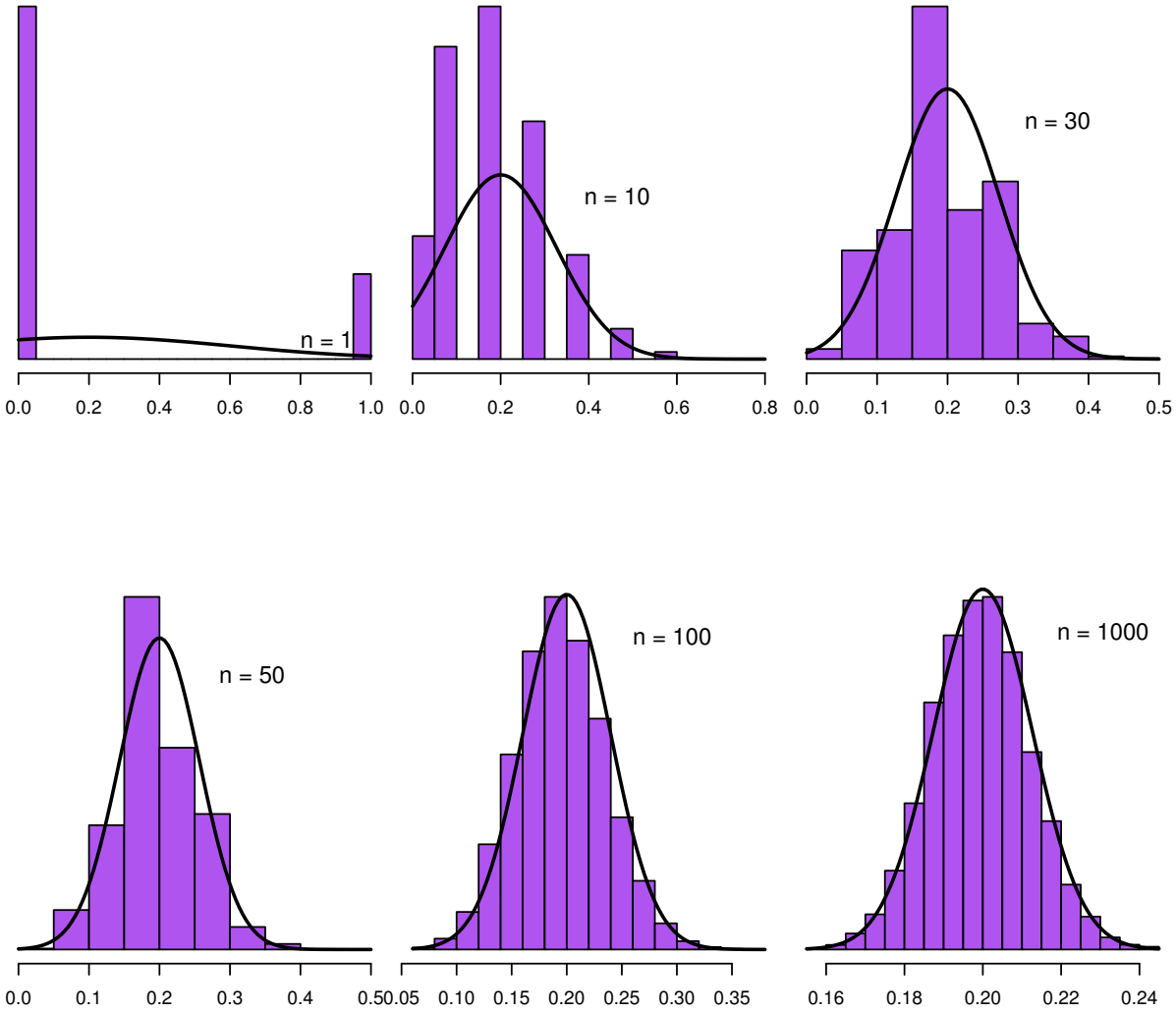
$$n = 1, 10, 30, 50, 100 \text{ and } 1000$$

from the Bernoulli(0.2) distribution (for which $\mu = 0.2$ and $\sigma^2 = 0.16$).

Here the distribution of X_i itself is not even continuous, and has only two possible values, 0 and 1. Nevertheless, the sampling distribution of $\bar{X}$ can be very well-approximated by the normal distribution, when n is large enough.

Note that since here $X_i = 1$ or $X_i = 0$ for all i,$\bar{X} = \sum_{i=1}^{n} X_i/n = m/n$, where m is the number of observations for which $X_i = 1$. In other words, $\bar{X}$ is the **sample proportion** of the value $X = 1$.

The normal approximation is clearly very bad for small n, but reasonably good already for $n = 50$, as shown by the histograms below.



Note that as n increases:

- there is convergence to $N(\mu, \sigma^2/n)$
- the sampling variance decreases (although the histograms might at first seems to show the same variation, look closely at the scale on the x-axes).

Sampling distribution of the sample proportion

The above example considered Bernoulli sampling where we noted that the sample mean was the sample proportion of successes, which we now denote as P.

Since from the CLT:

$$\bar{\boldsymbol{X}} \sim \boldsymbol{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\sigma^2}}{\boldsymbol{n}}\right)$$

approximately, when n is sufficiently large, and noting that when $X \sim \text{Bernoulli}(\pi)$ then:

$$\mathbf{E}(\boldsymbol{X}) = \mathbf{0} \times (\mathbf{1} - \boldsymbol{\pi}) + \mathbf{1} \times \boldsymbol{\pi} = \boldsymbol{\pi} = \boldsymbol{\mu}$$

and:

$$\mathbf{Var}(\boldsymbol{X}) = \boldsymbol{\pi}\,(\mathbf{1} - \boldsymbol{\pi}) = \boldsymbol{\sigma}^{\mathbf{2}}$$

we have:

$$\bar{\boldsymbol{X}} = \boldsymbol{P} \to \boldsymbol{N}\left(\boldsymbol{\mu},\, \frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\boldsymbol{n}}\right) = \boldsymbol{N}\left(\boldsymbol{\pi},\, \frac{\boldsymbol{\pi}\,(\mathbf{1} - \boldsymbol{\pi})}{\boldsymbol{n}}\right)$$

as $n \to \infty$.

We see that:

$$\mathbf{E}(\bar{\boldsymbol{X}}) = \mathbf{E}(\boldsymbol{P}) = \boldsymbol{\pi}$$

hence the sample proportion is equal to the population proportion, *on average*.[1] Also:

$$\mathbf{Var}(\boldsymbol{P}) \to \mathbf{0} \quad \textbf{as } \boldsymbol{n} \to \boldsymbol{\infty}$$

so the sampling variance tends to zero as the sample size tends to infinity, as we see in the histograms in the previous example.

[1] This means P is an unbiased estimator of π.