The central limit theorem (CLT)

Suppose that $Y_1,\ldots,Y_n \sim iid\ (\mu,\sigma^2)$, where $-\infty<\mu<\infty$ and $0<\sigma^2<\infty$.

Let $U_n=\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$. Then $U_n \xrightarrow{d} N(0,1)$ as $n\to\infty$.

Notes: The CLT makes no particular distributional assumptions about the Y_i values.

It assumes only that these are iid (independently and identically distributed) random variables which are not necessarily normal (NB: "*N*" is *not* missing in "(μ,σ^2)"), and with common finite mean μ and common finite (and non-zero) variance σ^2.

The CLT states that U_n *converges in distribution* to the standard normal distribution.

This means that it's cdf converges to the standard normal cdf as *n* tends to infinity, i.e.

$$F_{U_n}(u)\to F_{N(0,1)}(u),$$

or equivalently,

$$P(U_n\le u)\to\int_{-\infty}^{u}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}\,dt$$ (for all values *u* in the whole real line).

The proof of the CLT is beyond the scope of this course but can be achieved using the mgf technique, as in Section 7.4 (which is non-assessable).

The CLT implies that when *n* is 'large', it is reasonable to make the following approximation: $\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\overset{\sim}{\sim} N(0,1)$.

Example 5 200 numbers are randomly chosen from between 0 and 1. Find the probability that the average of these numbers is greater than 0.53.

Let Y_i be the *i*th number, *i* = 1,..., *n*, where *n* = 200. Then $Y_1,\ldots,Y_n$ ~ iid *U*(0,1).
Thus $\mu=EY_i=1/2$ and $\sigma^2=VarY_i$ = 1/12.
So we may also write $Y_1,\ldots,Y_n \sim iid\ (\mu,\sigma^2)$. Applying the CLT, we find that

$$P(\bar{Y}>0.53)=P\left(\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}>\frac{0.53-1/2}{\sqrt{1/12}/\sqrt{200}}\right)\approx P(Z>1.47)\quad \text{where } Z\sim N(0,1)$$

= 0.0708 using normal tables.

Since 200 is large, 7.08% will be close to the true probability, which would be very difficult (or virtually impossible) to work out exactly.

Another way to think about the CLT is: $\bar{Y} \dot{\sim} N(\mu, \sigma^2/n)$.

Thus in Example 5, $\bar{Y} \dot{\sim} N(1/2, (1/12)/200)$.
So $P(\bar{Y} > 0.53) \approx P(U > 0.53)$ where $U \sim N(1/2, 1/2400)$

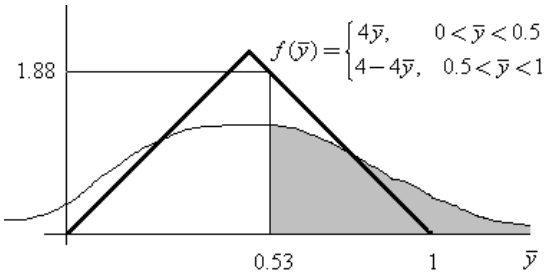
$$= P\left(Z > \frac{0.53 - 1/2}{\sqrt{1/2400}}\right) = 0.0708.$$

A good exercise related to Example 5 is to draw 4 graphs, one for each n = 1,2,3,4. In each graph, derive and draw the *exact* pdf of $\bar{Y}$ and also the *approximate* pdf of $\bar{Y}$ based on the CLT, namely the pdf of $U \sim N(1/2, 1/(12n))$. Then calculate the exact and approximate values of $P(\bar{Y} > 0.53)$. Illustrate both probabilities in each graph.

You will find that the exact pdf of $\bar{Y}$ is:

- flat (or constant) when n = 1
- triangular when n = 2 (see below)
- made up of 3 quadratics (joined at 1/3 and 2/3) when n = 3
- made up of 4 cubics (joined at 1/4, 1/2 and 3/4) when n = 4, etc.

For example, consider n = 2. The following is the required figure (roughly).



In this case, $P(\bar{Y} > 0.53) = \frac{0.47 \times 1.88}{2} = 0.44$ (exact).

(This is the area of the triangle defined by (0.53,0), (1,0) and (0.53,1.88).)

Let $U \sim N(1/2, 1/24)$ (the normal dsn with the same mean and variance as $\bar{Y}$).

Then $P(\bar{Y} > 0.53) \approx P(U > 0.53) = P\left(Z > \frac{0.53 - 1/2}{\sqrt{1/24}}\right) = P(Z > 0.15) = 0.4404.$

(This is the area of the shaded portion in the figure.)

NB: The approximation 0.4404 is incredibly close to the exact probability 0.44 here, even though n is very small (2). But this is just luck, and we usually need n to be at least about 20 to feel fairly confident that the CLT is providing a good approximation.

Yet another way to think about the CLT is: $\dot{Y} \dot{\sim} N(n\mu, n\sigma^2)$,
where $\dot{Y} = Y_1 + \ldots + Y_n$ (the *sample total*).

Example 6 A die is about to be rolled 50 times, and each time you will win as many dollars as the number which comes up.

Find the probability that you will win a total of at least \$200.

Let Y_i be the number of dollars you will win on the *i*th roll.

Then $Y_1, \ldots, Y_n \sim iid\ (\mu, \sigma^2)$,
where: $\mu = EY_i = 3.5$
$\sigma^2 = VarY_i$ = 2.9167 (see Tutorial 5, Problem 1).

Therefore $P(\dot{Y} \geq 200) \approx P(U > 200)$, where $U \sim N(50(3.5), 50(2.9167))$
= $N(175, 145.83)$

$$= P\left(Z > \frac{200-175}{\sqrt{145.83}}\right)$$

$$= P(Z > 2.07)$$

$$= 0.0192.$$

The normal approximation to the binomial distribution

Suppose that $Y \sim Bin(n,p)$.
Then $Y = Y_1 + \ldots + Y_n$, where $Y_1, \ldots, Y_n \sim iid\ Bern(p)$.
So $Y_1, \ldots, Y_n \sim$ iid (μ, σ^2), where: $\mu = EY_i = p$
$\sigma^2 = VarY_i = p(1-p)$.
It follows by the CLT that $Y \dot{\sim} N(np, np(1-p))$.

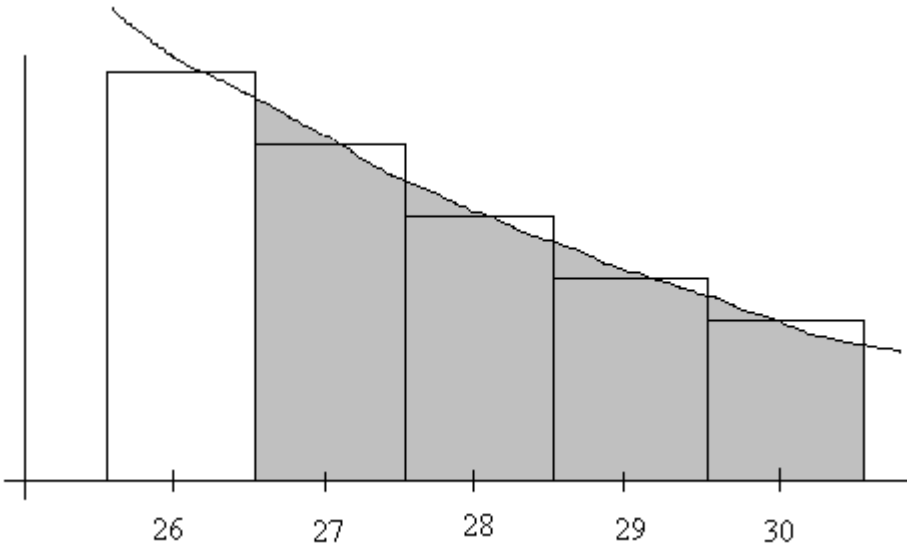
Example 7 A die is rolled n = 120 times.
Find the probability that at least 27 sixes come up.

Let Y be the number of 6's. Then $Y \sim Bin(120,1/6)$.
So $Y \dot{\sim} N\left(120\left(\frac{1}{6}\right), 120\left(\frac{1}{6}\right)\left(1-\frac{1}{6}\right)\right)$.
Therefore $P(Y \geq 27) \approx P(U > 27)$ where $U \sim N(20,16.667)$

$$= P\left(Z > \frac{27-20}{\sqrt{16.667}}\right) = P(Z > 1.71) = 0.0436.$$

The continuity correction

Let's take a closer look at the approximation made in the last example.



The exact probability is the area of the boxes above 27, 28, 29, ...

We have approximated this probability by **<u>0.0436</u>**, which is the area under the approximating normal density to the right of 27.0. But this area seems to be too small by about half the area of the box above 27 (i.e. the left half of that box).

Thus, it would appear that a better approximation is the area to the right of 26.5 = 27 – 0.5 (shaded above). We call "–0.5" here the *continuity correction*.

Let's now apply this continuity correction and see whether it makes much difference.

$$P(Y \geq 27) \approx P(U > 27 - 0.5) \quad \text{where } U \sim N(20, 16.667)$$

$$= P\left(Z > \frac{27 - 0.5 - 20}{\sqrt{16.667}}\right) = P(Z > 1.59) = \mathbf{\underline{0.0559}}.$$

Now, the exact probability can be calculated using a computer, and it works out as

$$P(Y \geq 27) = \sum_{y=27}^{120} \binom{120}{y} \left(\frac{1}{6}\right)^{y} \left(\frac{5}{6}\right)^{120-y} = \mathbf{\underline{0.0597}}.$$

We see that the continuity correction here does indeed improve the approximation. This is typically the case. However, note that sometimes the correction makes very little difference (e.g. if n is very large), and sometimes it produces a worse answer.