

10 The Central Limit Theorem

The sum of two random variables

Say we have two continuous random variables X and Y , with density functions $p_X(x)$ and $p_Y(y)$; consider working out the distribution for

$$Z = X + Y \quad (1)$$

Now the thing is that there is more than one value for x and y that gives $Z = z$. In fact $Z = z$ if $Y = z - x$ for any value of x so

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)dx \quad (2)$$

This calculation is called a *convolution*.

As a simple example consider X and Y both uniformly distributed over $[0, 1]$:

$$p_X(x) = \begin{cases} 1 & : x \in [0, 1] \\ 0 & : \text{otherwise} \end{cases} \quad (3)$$

with $p_Y(y)$ the same. Now $Z = X + Y$ has the distribution

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x)p_Y(z-x)dx \quad (4)$$

where the integrand is zero unless x and $z - x$ are both inside the interval zero to one. Now, the first constraint, $x \in [0, 1]$ is easy but the second one depends on z . If $z < 0$ then $z - x$ is always negative and $p_Y(z - x) = 0$. If $z > 2$ then $z - x > 1$ and again $p_Y(z - x) = 0$. If $z \in [0, 2]$ there are two cases, if $z < 1$ then $z > 0$ for $x \in [0, z]$, otherwise if $z > 1$ then $z < 1$ for $x \in [1 - z, 1]$. Lets do the two cases separately, for $z \in [0, 1]$:

$$p_Z(z) = \int_0^z dx = z \quad (5)$$

and for $z \in (1, 2]$

$$p_Z(z) = \int_{1-z}^1 dx = 1 - (1 - z) = -z \quad (6)$$

Putting it all together we get

$$p_Z(z) = \begin{cases} z & z \in [0, 1] \\ -z & z \in [1, 2] \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Sums and the moment generating function¹

Recall that

$$m_X(t) = \langle e^{tX} \rangle \quad (8)$$

where we have used the usual trick of putting a subscript X for distinguishing probability quantities for different random variables. Now if X and Y are independent we have

$$m_{X+Y}(t) = \langle e^{t(X+Y)} \rangle = \langle e^{tX} e^{tY} \rangle = \langle e^{tX} \rangle \langle e^{tY} \rangle = m_X(t)m_Y(t) \quad (9)$$

¹not covered in lectures and not examinable

The sum of two Gaussians²

Next let's consider the sum of two Gaussians, $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$. Recall that we know the moment generating map for the Gaussian and

$$\begin{aligned}m_X(t) &= e^{\mu_x t + \frac{1}{2}\sigma_x^2 t^2} \\m_Y(t) &= e^{\mu_y t + \frac{1}{2}\sigma_y^2 t^2}\end{aligned}\tag{10}$$

so if $Z = X + Y$, using the formula above for the moment generating map of a sum

$$m_Z(t) = e^{\mu_x t + \frac{1}{2}\sigma_x^2 t^2} e^{\mu_y t + \frac{1}{2}\sigma_y^2 t^2} = e^{(\mu_x + \mu_y)t + \frac{1}{2}(\sigma_x^2 + \sigma_y^2)t^2}\tag{11}$$

Hence

$$Z \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)\tag{12}$$

The central limit theorem

Let $\{X_1, X_2, \dots, X_n\}$ be a set of random variables. A set of random variables is called independent identically distributed, usually abbreviated to i.i.d., if the variables all have the same probability density, say $p_X(x)$ and are independent. Let μ and σ^2 be the mean and variance of that probability density. Now consider the *sample mean*:

$$S_n = \frac{X_1 + X_2 + \dots + X_n}{n}\tag{13}$$

If you were estimating the mean for X this is the value you would measure if you took n independent samples. Now the central limit theorem states that for n approaching infinity

$$U_n = \sqrt{n} \left(\frac{S_n - \mu}{\sigma} \right) \sim \mathcal{N}(0, 1)\tag{14}$$

As an example, consider the uniform distribution example we looked before; in Fig. 1 we plot U_n for different n . We see that it very quickly comes to resemble a Gaussian distribution.

We won't prove the central limit theorem here, doing so involves a bit of detail in deciding what we mean by 'for n approaching infinity'. The proof itself involves some messing about with the moment generating theorem and isn't too hard. It isn't particularly revealing though; it misses what is happening, which is roughly this: the central limit theorem is a bit like the binomial distribution, just like the binomial distribution quantifies the different ways to get different numbers of S , the central limit theorem quantifies different ways of getting the same sum. In fact, numerically, the Gaussian distribution is close to the binomial distribution, see for example Fig. 2.

The central limit theorem is important in two ways. First off, it tells us how the sample mean behaves; since we often estimate means by doing multiple independent trials this is useful. The second point is vaguer, although we have stated the central limit theorem in terms of i.i.d. variables and this is the case where it is easy to state and prove the theorem, there is a more general phenomena, which is that if you add independent variables, even if they are not identical, then the result is a Gaussian. This is why so many things show Gaussian distributions. Think, for example, of the height of trees, lots of things contribute to the height

²not covered in lectures and not examinable

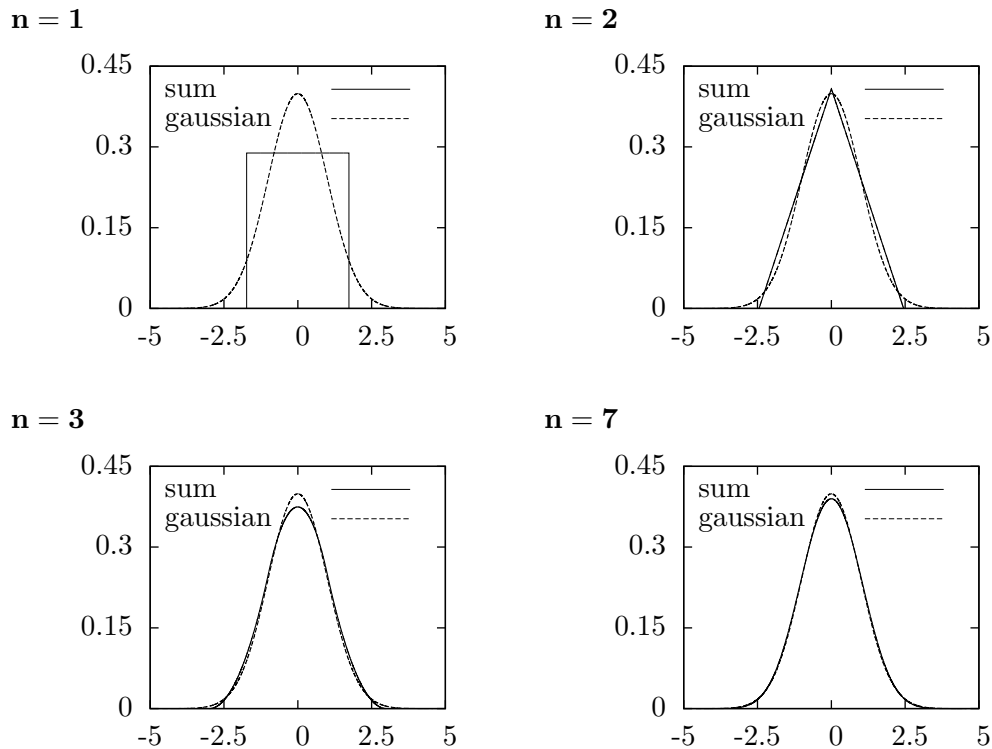


Figure 1: U_n for different n s where X is the uniform distribution.

of a tree, how wet and sunny the location it is growing in, the soil, the genetics of each individual. These individual contributions probably aren't Gaussian, they aren't identical and, in fact, probably don't contribute in a linear way to the height, nonetheless we would expect the result, the height of the trees to be Gaussian. It is a fortunate aspect of statistics, things that are complicated and have multiple contributing factors, tend to behave in a simple way from the point-of-view of statistics.

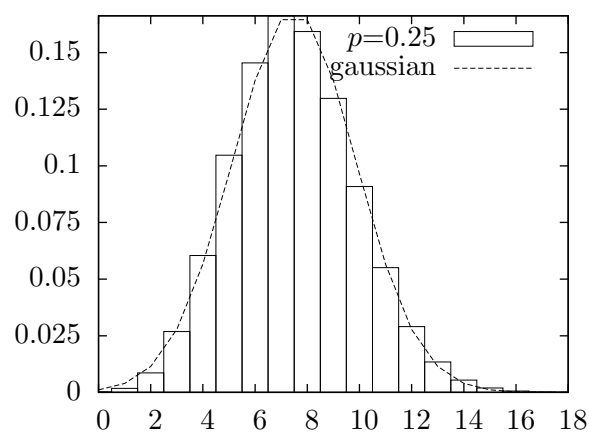


Figure 2: This compares the binomial distribution with $n = 30$ and $p = 0.25$ with the Gaussian distribution with the same mean and variance.