# 9 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

$$Z = X + Y \tag{1}$$

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

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

This calculation is called a **convolution**.

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

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

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

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

where the integrand is zero unless $x$ and $z - x$ are both inside the interval zero to one. Now, the first contraint, $x \in [0, 1]$ 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 $x$ then $z - x > 0$ for $x \in [0, z]$, otherwise if $z > 1$ then $z - x < 1$ in $x \in [z - 1, 1]$. Lets do the two cases separately, for $z \in [0, 1]$:

$$p_Z(z) = \int_0^z dx = z \tag{5}$$

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

$$p_Z(z) = \int_{z-1}^1 dx = 1 - (z - 1) = 2 - z \tag{6}$$

Putting it all together we get

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

## The sum of two Gaußians

Next lets consider the sum of two Gaußians, $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Calculating the distribution of $Z = X + Y$ would be difficult since it involves the integral

$$p_Z(z) = \frac{1}{2\pi\sigma_X^2\sigma_Y^2} \int_{-\infty}^{\infty} e^{-(x-\mu_X)^2/2\sigma_X^2 - (z-x-\mu_Y)^2/2\sigma_Y^2} dx \tag{8}$$

We aren't going to attempt that, but we will quote the result: the sum is also a Gaußian:

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \tag{9}$$

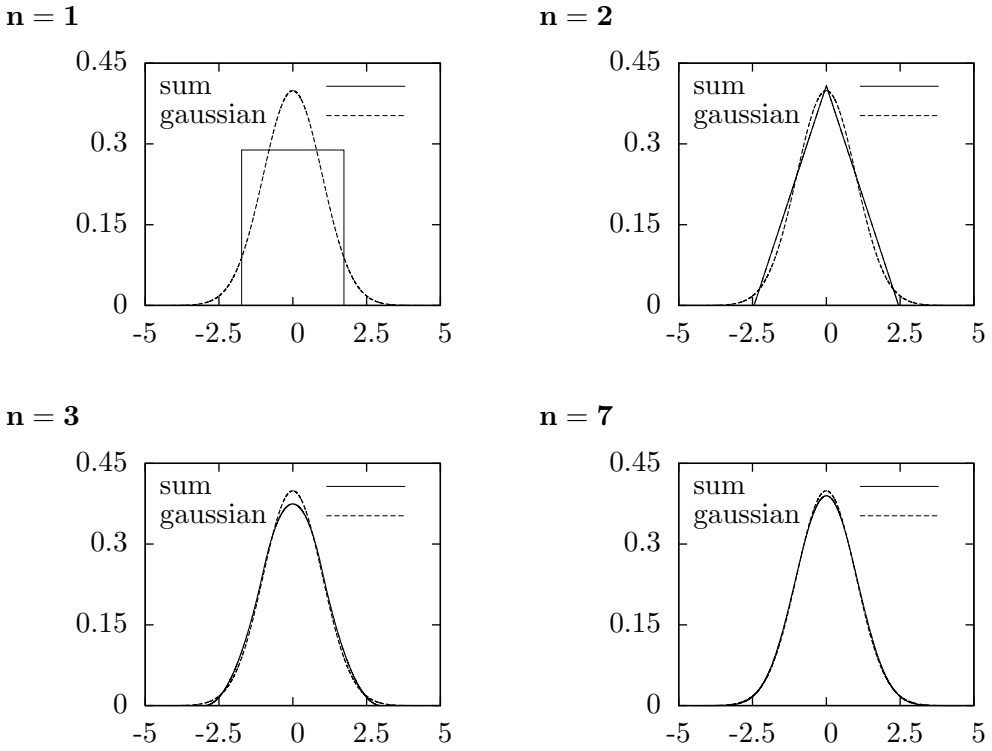


Figure 1: $U_n$ for different $n$s where $X$ is the uniform distribution.

## The central limit theorem

Let $\{X_1, X_2, \ldots, X_n\}$ 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 $\mu$ and $\sigma^2$ be the mean and variance of that probability density. Now consider the **sample mean**:

$$S_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \tag{10}$$

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 stats that for $n$ approaching infinity

$$U_n = \sqrt{n}\left(\frac{S_n - \mu}{\sigma}\right) \sim \mathcal{N}(0, 1) \tag{11}$$

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 which we haven't studied, but otherwise isn't too hard. It isn't particularly revealing though; it misses what is happening, which is roughly this: the

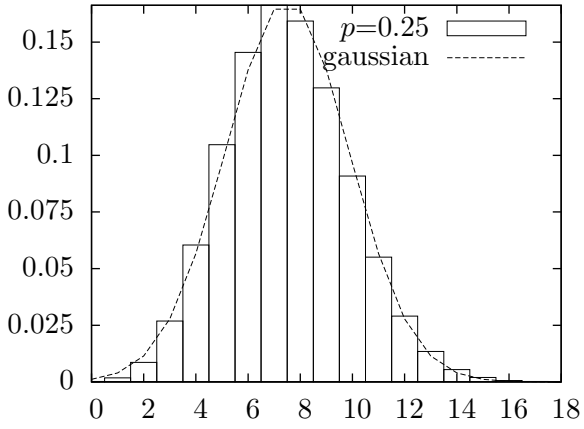


Figure 2: This compares the binomial distribution with $n = 30$ and $p = 0.25$ with the Gaußian distribution with the same mean and variance.

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 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.

## Summary

- If $X$ and $Y$ are continuous random variables, with density functions $p_X(x)$ and $p_Y(y)$ and

$$Z = X + Y \tag{12}$$

then

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x) p_Y(z - x) dx \tag{13}$$

This calculation is called a **convolution**.

- If $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ then

$$X + Y = Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \tag{14}$$

- Let $\{X_1, X_2, \ldots, X_n\}$ be a set of random variables. A set of random variables is called i**ndependent identically distributed**, usually abbreviated to i.i.d., if the variables all have the same probability density, say $p_X(x)$ and are independent.

- The **Central limit theorem**: if $\{X_1, X_2, \ldots, X_n\}$ is i.i,d, the **sample mean** is

$$S_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \tag{15}$$

As $n$ approaches infinity

$$U_n = \sqrt{n} \left( \frac{S_n - \mu}{\sigma} \right) \sim \mathcal{N}(0, 1) \tag{16}$$