

Sums of independent random variables

Theorems about the distribution of a sum

If $X_1, X_2, \dots, X_m$ are independent random variables such that $X_i \sim \text{Bin}(n_i, p)$ for $i = 1, 2, \dots, m$, and $Y = X_1 + X_2 + \dots + X_m$, then Theorem 5 tells us that $Y \sim \text{Bin}(N, p)$ where $N = n_1 + n_2 + \dots + n_m$.

Similarly, if $X_1, X_2, \dots, X_m$ are independent random variables for which $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 1, 2, \dots, m$, and $Y = X_1 + X_2 + \dots + X_m$, then Theorem 6 tells us that $Y \sim \text{Poisson}(\mu)$ where $\mu = \lambda_1 + \lambda_2 + \dots + \lambda_m$.

Theorem 15 If $X_1, X_2, \dots, X_n$ are mutually independent normal random variables, and $a_0, a_1, \dots, a_n$ are any real numbers, and

$$Y = a_0 + \sum_{i=1}^n a_i X_i,$$

then Y is normal.

The proof is not given in this module, but we can easily find the expectation and variance of Y . Suppose that $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \dots, n$. Then

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}\left(a_0 + \sum_{i=1}^n a_i X_i\right) \\ &= \mathbb{E}(a_0) + \sum_{i=1}^n \mathbb{E}(a_i X_i) \\ &= a_0 + \sum_{i=1}^n a_i \mathbb{E}(X_i) \\ &= a_0 + \sum_{i=1}^n a_i \mu_i. \end{aligned}$$

Moreover,

$$\begin{aligned}
\text{Var}(Y) &= \text{Var}\left(a_0 + \sum_{i=1}^n a_i X_i\right) \\
&= \text{Var}\left(\sum_{i=1}^n a_i X_i\right) \\
&= \sum_{i=1}^n \text{Var}(a_i X_i), \quad \text{since } X_1, X_2, \dots, X_n \text{ are mutually independent,} \\
&= \sum_{i=1}^n a_i^2 \text{Var}(X_i) \\
&= \sum_{i=1}^n a_i^2 \sigma_i^2.
\end{aligned}$$

The proofs of the following two theorems will also be given in later modules.

Theorem 16 (The Law of Large Numbers) Let $X_1, X_2, \dots$ be a sequence of mutually independent random variables which have identical distributions with finite mean μ and finite variance. For $n \geq 1$, define the partial sum S_n by

$$S_n = X_1 + X_2 + \dots + X_n$$

and the partial mean $\bar{X}_n$ by

$$\bar{X}_n = \frac{S_n}{n}.$$

Then $\bar{X}_n$ converges in distribution to μ as n tends to infinity.

This means that the cumulative distribution function of $\bar{X}_n$ converges to the cumulative distribution function for the constant μ . We can write this as follows:

$$\mathbb{P}(\bar{X}_n \leq x) \rightarrow \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x \geq \mu. \end{cases}$$

Theorem 17 (The Central Limit Theorem) Let $X_1, X_2, \dots$ be a sequence of mutually independent random variables which have finite means and variances

$$\mathbb{E}(X_i) = \mu_i \quad \text{and} \quad \text{Var}(X_i) = \sigma_i^2$$

for $i = 1, 2, \dots$. Put

$$S_n = \sum_{i=1}^n X_i$$

and

$$Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}.$$

Then Z_n converges in distribution to the standard normal $N(0, 1)$ as $n \rightarrow \infty$.

Another way of saying this is that if F_n is the cumulative distribution function of Z_n and Φ is the cumulative distribution function of the standard normal distribution then

$$\lim_{n \rightarrow \infty} \frac{F_n(x)}{\Phi(x)} = 1 \quad \text{for } x \text{ in } \mathbb{R}.$$

We can write it in this way because $\Phi(x) \neq 0$ for all x in $\mathbb{R}$.

The most useful special case of this occurs when $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$ for all i . Then

$$Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\frac{S_n}{n} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}.$$

Corollary Let $X_1, X_2, \dots$ be a sequence of mutually independent identically distributed random variables with finite mean μ and finite variance σ^2 . Then the distribution of

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

tends to the $N(0, 1)$ distribution as $n \rightarrow \infty$.

Normal approximations

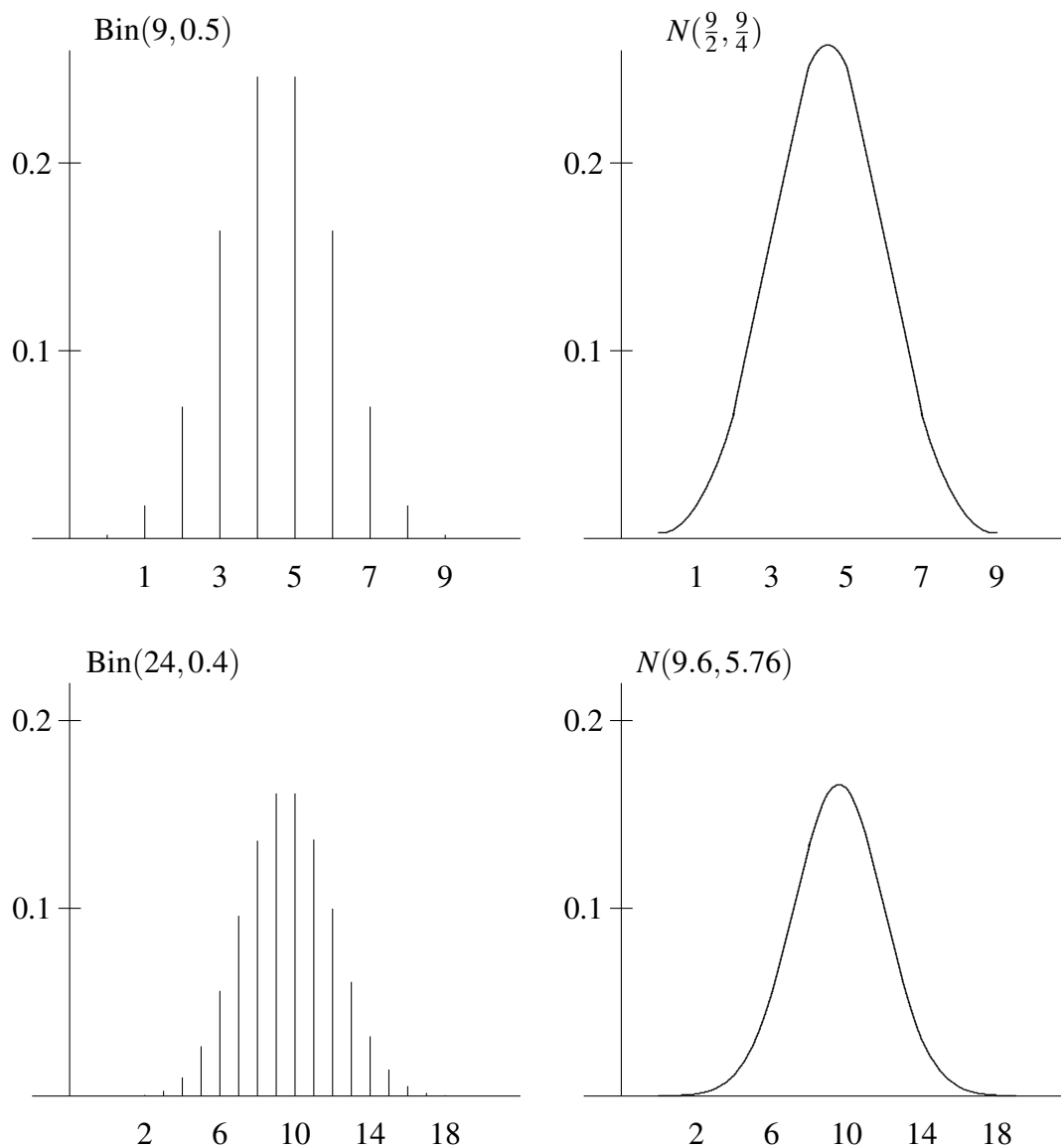
The Central Limit Theorem tells us that the normal distribution is a good approximation to many random variables which arise naturally, especially those which are sums of many random parts. Examples include

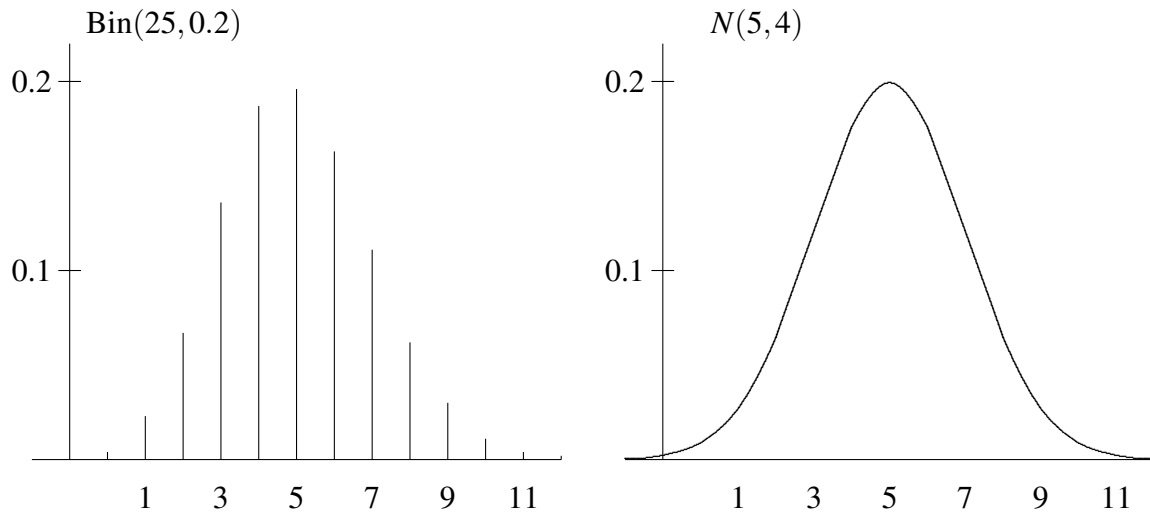
- people's heights in cm (these must be positive, so if the distribution is approximately $N(\mu, \sigma^2)$ then we must have $\mu - 3\sigma > 0$);
- yields of wheat in tonnes per hectare.

If $X \sim \text{Poisson}(\lambda)$ then X is a sum of 100 independent random variables with distribution $\text{Poisson}(\lambda/100)$, so the Central Limit Theorem suggests that X is approximately normal. But $\mathbb{E}(X) = \lambda$ and $\text{Var}(X) = \lambda$, so this cannot be a good approximation unless $\lambda - 3\sqrt{\lambda} > 0$; that is, $\lambda > 9$.

If $X \sim \text{Bin}(n, p)$ then X is a sum of n independent random variables with distribution $\text{Bin}(1, p) = \text{Bernoulli}(p)$, so the Central Limit Theorem shows that X is approximately $N(np, npq)$ when n is large, where $q = 1 - p$. This cannot be a good approximation unless $np > 3\sqrt{npq}$, that is $n^2 p^2 > 9npq$, that is $np > 9q$. If X is approximately normal then $n - X$ must also be approximately normal, but this is $\text{Bin}(n, q)$, so the same argument gives $nq > 9p$. The larger the difference between p and q , the larger n needs to be before the approximation is good.

The following graphs show three examples: the less symmetric is the original distribution, the larger that n needs to be before the approximation is good.





If p is very small then $q \approx 1$, so if $X \sim \text{Bin}(n, p)$ then $\mathbb{E}(X) \approx \text{Var}(X)$. Thus the distribution $\text{Poisson}(np)$ is a good approximation to $\text{Bin}(n, p)$ for smaller values of n than the normal distribution is.

In Practical 9, you drew some graphs to show how well the normal distribution approximates the binomial and Poisson distributions. This would be a good place for you to insert those graphs into your notes.

The continuity correction

If X is a *discrete* random variable with integer values then we must use the *continuity correction* if we approximate it by a continuous distribution like the normal. Since X takes only integer values, we have $\mathbb{P}(X \leq 79) + \mathbb{P}(X \geq 80) = 1$, but if we approximate X by a continuous random variable Y then there is a positive probability that Y takes values between 79 and 80. To make sure that our probabilities still add up to 1, we approximate $\mathbb{P}(X \leq 79)$ by $\mathbb{P}(Y \leq 79.5)$ and $\mathbb{P}(X \geq 80)$ by $\mathbb{P}(Y \geq 79.5)$.

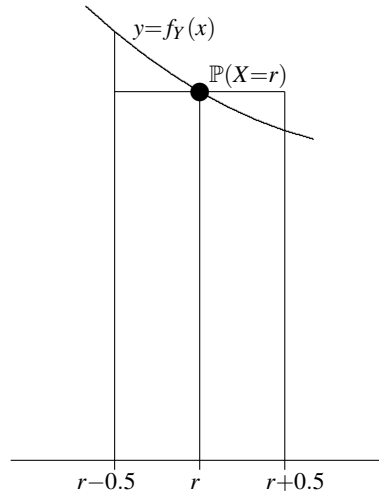
In general, suppose that X takes integer values, with $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$. In this module we consider approximating X by the continuous random variable Y , where $Y \sim N(\mu, \sigma^2)$. If r and s are integers with $r < s$, then we use the approximation

$$\mathbb{P}(r \leq X \leq s) = \mathbb{P}(r - 0.5 \leq X \leq s + 0.5) \approx \mathbb{P}(r - 0.5 \leq Y \leq s + 0.5).$$

Here is another way of seeing this. The probability we want is equal to

$$\mathbb{P}(X = r) + \mathbb{P}(X = r + 1) + \cdots + \mathbb{P}(X = s).$$

To say that X can be approximated by Y means that $\mathbb{P}(X = r)$ is approximately equal to $f_Y(r)$, where f_Y is the probability density function of Y . This is equal to the area of a rectangle of height $f_Y(r)$ and base 1 (from $r - 0.5$ to $r + 0.5$). This in turn is, to a good approximation, the area under the curve $y = f_Y(x)$ from $x = r - 0.5$ to $x = r + 0.5$, since the pieces of the curve above and below the rectangle on either side of $x = r$ will approximately cancel. Similarly for the other values.



Adding all these pieces, we find that $\mathbb{P}(r \leq X \leq s)$ is approximately equal to the area under the curve $y = f_Y(x)$ from $x = r - 0.5$ to $x = s + 0.5$. This area is given by $F_Y(s + 0.5) - F_Y(r - 0.5)$, since F_Y is the integral of f_Y . Said otherwise, this is $\mathbb{P}(r - 0.5 \leq Y \leq s + 0.5)$.

Example A fair coin is tossed 100 times. Let X be the number of heads. Then $X \sim \text{Bin}(100, 0.5)$, with mean $= 100 \times 0.5 = 50$ and variance $= 100 \times 0.5 \times (1 - 0.5) = 25$. So X is approximately distributed like Y , where $Y \sim N(50, 25)$. Therefore

$$\begin{aligned}
 \mathbb{P}(X \leq 57) &= \mathbb{P}(X \leq 57.5) \\
 &\approx \mathbb{P}(Y \leq 57.5) \\
 &= \mathbb{P}\left(\frac{Y - 50}{5} \leq \frac{57.5 - 50}{5}\right) \\
 &= \mathbb{P}(Z \leq 1.5), \quad \text{where } Z \sim N(0, 1), \\
 &= \Phi(1.5) \\
 &= 0.9332,
 \end{aligned}$$

from Table 4 of the *New Cambridge Statistical Tables* [1].

Example The probability that a light bulb will fail in a year is 0.75, and light bulbs fail independently. If 192 bulbs are installed, what is the probability that the number which fail in a year lies between 140 and 150 inclusive?

Let X be the number of light bulbs which fail in a year. Then $X \sim \text{Bin}(192, 3/4)$, and so $\mathbb{E}(X) = 144$ and $\text{Var}(X) = 36$. Therefore X is approximated by Y , where $Y \sim N(144, 36)$, and

$$\begin{aligned}
 \mathbb{P}(140 \leq X \leq 150) &= \mathbb{P}(139.5 \leq X \leq 150.5) \\
 &\approx \mathbb{P}(139.5 \leq Y \leq 150.5),
 \end{aligned}$$

by the continuity correction.

Let $Z = (Y - 144)/6$. Then $Z \sim N(0, 1)$, and

$$\begin{aligned}
 \mathbb{P}(139.5 \leq Y \leq 150.5) &= \mathbb{P}\left(\frac{139.5 - 144}{6} \leq Z \leq \frac{150.5 - 144}{6}\right) \\
 &= \mathbb{P}(-0.75 \leq Z \leq 1.083) \\
 &= \Phi(1.083) - \Phi(-0.75) \\
 &= 0.8606 - 0.2266 \quad (\text{from Table 4 of NCST}) \\
 &= 0.6340.
 \end{aligned}$$

Example It was reported in the local newspaper that, out of 200 local schools, 80 came in the top third of the national league table for a certain test. Is this especially praiseworthy?

Let X be the number in the top third if the 200 local schools are a random sample of all schools in the country. Then

$$X \sim \text{Hg}(200, \frac{N}{3}, N),$$

where N is the number of schools in the country. But N is much larger than 200, so approximately

$$X \sim \text{Bin}(200, \frac{1}{3}).$$

Therefore

$$\mathbb{E}(X) = \frac{200}{3} \quad \text{and} \quad \text{Var}(X) = 200 \times \frac{1}{3} \times \frac{2}{3} = \frac{400}{9},$$

and so X is approximately distributed like Y , where

$$Y \sim N\left(\frac{200}{3}, \frac{400}{9}\right).$$

Therefore

$$\begin{aligned} \mathbb{P}(X \geq 80) &= \mathbb{P}(X \geq 79\frac{1}{2}), && \text{by the continuity correction,} \\ &\approx \mathbb{P}(Y \geq 79\frac{1}{2}) \\ &= \mathbb{P}\left(\frac{Y - 200/3}{20/3} \geq \frac{79\frac{1}{2} - 200/3}{20/3}\right) \\ &= \mathbb{P}\left(Z \geq \frac{238\frac{1}{2} - 200}{20}\right), && \text{where } Z \sim N(0, 1), \\ &= \mathbb{P}(Z \geq 1.925) \\ &= 1 - \Phi(1.925) \\ &= 1 - 0.9729, && \text{using interpolation,} \\ &= 0.0271. \end{aligned}$$

[1] D. V. Lindley and W. F. Scott, *New Cambridge Statistical Tables*, Cambridge University Press.