

## The Central Limit Theorem

A more formal and mathematical statement of the Central Limit Theorem is stated in the following way.

**The Central Limit Theorem:** Suppose that  $Y_1, Y_2, \dots, Y_n$  are independent and identically distributed with mean  $\mu$  and finite variance  $\sigma^2$ . Define the random variable  $U_n$  as follows:

$$U_n = \frac{(\bar{Y} - \mu)}{\left(\frac{\sigma}{\sqrt{n}}\right)}, \text{ where } \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

Then the distribution function of  $U_n$  converges to the standard normal distribution function as  $n$  increases without bound.

*Proof:*

Define a random variable  $Z_i$  by

$$Z_i = \frac{Y_i - \mu}{\sigma}.$$

Notice that  $E(Z_i) = 0$  and  $V(Z_i) = 1$ . Thus, the moment generating function can be written as

$$m_Z(t) = 1 + \frac{t^2}{2} + \frac{t^3}{3!} E(Z_i^3) + \dots.$$

Also, we know that

$$U_n = \sqrt{n} \left( \frac{\bar{Y} - \mu}{\sigma} \right) = \frac{1}{\sqrt{n}} \left( \frac{\sum_{i=1}^n Y_i - n\mu}{\sigma} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i.$$

Because the random variables  $Y_i$  are independent, so are the random variables  $Z_i$ . We know that the moment-generating function of the sum of independent random variables is the product of their individual moment-generating functions. Thus,

$$m_n(t) = \left[ m_Z \left( \frac{t}{\sqrt{n}} \right) \right]^n = \left( 1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}} E(Z_i^3) + \dots \right)^n$$

We now take the limit of  $m_n(t)$  as  $n \rightarrow \infty$ . This can be facilitated by considering

$$\ln(m_n(t)) = n \ln \left( 1 + \left( \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}} E(Z_i^3) + \dots \right) \right).$$

We now make a substitution into the expression on the right. Recall that the Taylor series expansion for  $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$ . If we let  $x = \left( \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}} E(Z_i^3) + \dots \right)$ , then

$$\ln(m_n(t)) = n \ln(1+x) = n \left( x - \frac{x^2}{2} + \frac{x^3}{3} - \dots \right).$$

Now, rewrite this last expression by substituting for  $x$ . This gives the very messy equation

$$\ln(m_n(t)) = n \left[ \left( \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}} E(Z_i^3) + \dots \right) - \frac{1}{2} \left( \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}} E(Z_i^3) + \dots \right)^2 + \frac{1}{3} \left( \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}} E(Z_i^3) + \dots \right)^3 - \dots \right]$$

If we multiply through by the initial  $n$ , all terms except the first have some positive power of  $n$  in the denominator. Consequently, as  $n \rightarrow \infty$ , all terms but the first go to zero, leaving

$$\lim_{n \rightarrow \infty} \ln(m_n(t)) = \frac{t^2}{2}$$

and

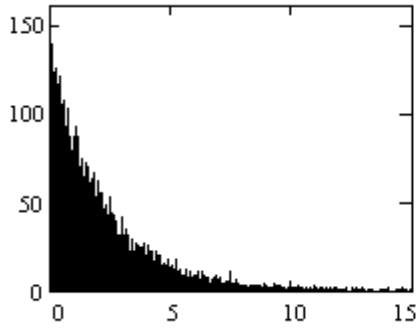
$$\lim_{n \rightarrow \infty} (m_n(t)) = e^{\frac{t^2}{2}}.$$

The last is recognized as the moment-generating function for a standard normal random variable. Since moment-generating functions are unique, and invoking the Theorem 2 above, we know that  $U_n$  has a distribution that converges to the distribution function of the standard normal random variable.

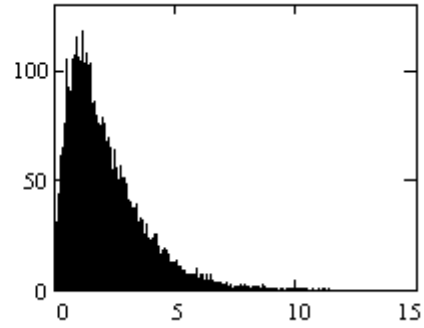
The essential implication here is that probability statements about  $U_n$  can be approximated by probability statements about the standard normal random variable if  $n$  is large.

## Simulations

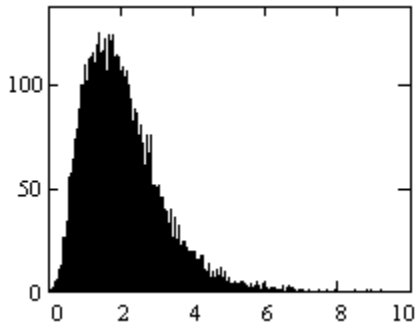
One way to "see" the Central Limit Theorem in action is through simulations. We can select  $n$  independent values from a given distribution, find the mean, and repeat the process several thousand times. Gathering all of the computed sample means, we can find the mean and standard deviation of these sample means and present the distribution in a histogram. A few examples are shown below. The population is  $\chi^2(2)$  which has  $\mu = 2$  and  $\sigma^2 = 4$ . If we draw  $n$  values from this distribution 25,000 times, compute the mean and standard deviation of these 25,000 draws, and plot a histogram of the results, we have the following graphs and values. We are expecting the mean and standard deviation of the 25,000 draws to be  $\bar{x} = 2$  and  $s = \frac{2}{\sqrt{n}}$ , respectively.



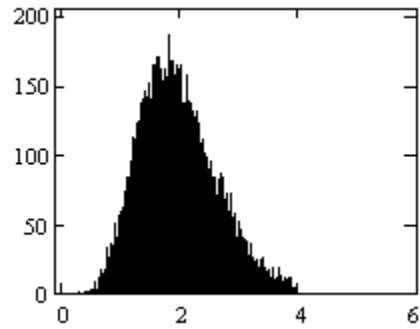
$n = 1, \bar{x} = 1.995, s = 2.007$



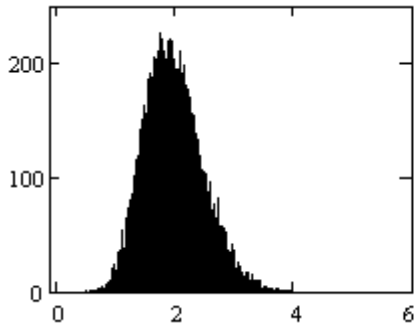
$n = 2, \bar{x} = 2.017, s = 1.434$



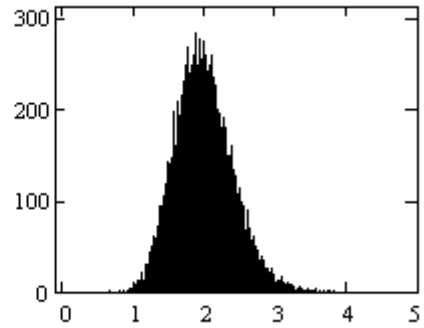
$n = 4, \bar{x} = 2.001, s = 1.005$



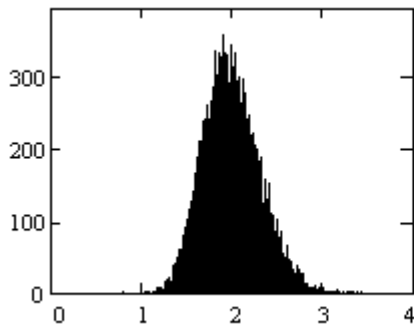
$n = 9, \bar{x} = 1.995, s = 0.664$



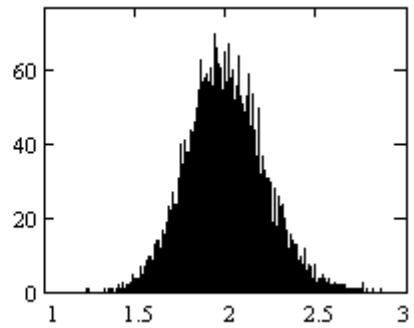
$n = 16, \bar{x} = 1.993, s = 0.469$



$n = 25, \bar{x} = 2.000, s = 0.401$

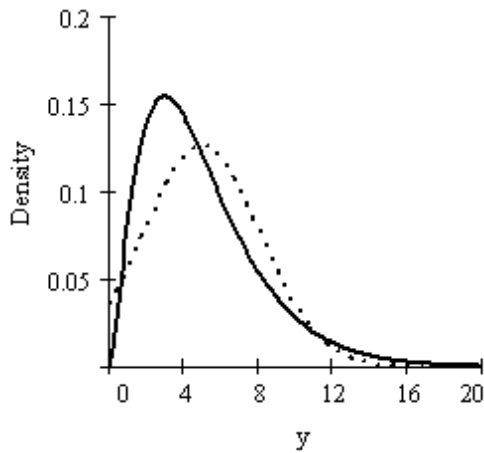


$n = 40, \bar{x} = 2.001, s = 0.315$

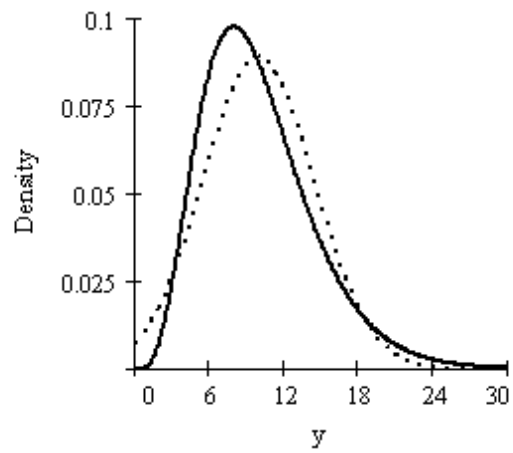


$n = 100, \bar{x} = 2.001, s = 0.200$

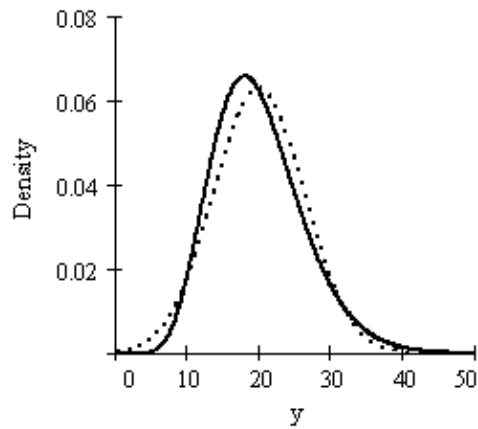
Another approach to visualizing the Central Limit Theorem is to compare the probability density functions. We expect that the distribution function for the chi-square distribution with  $\nu$  degrees of freedom will approach the normal distribution function with mean  $\nu$  and variance  $2\nu$ , that is,  $\chi^2(\nu) \approx N(\nu, 2\nu)$  and  $\nu$  increases. The graphs of these probability density functions are given below. In the first set of graphs, the chi-square probability density function is in bold and the normal density function is dashed.



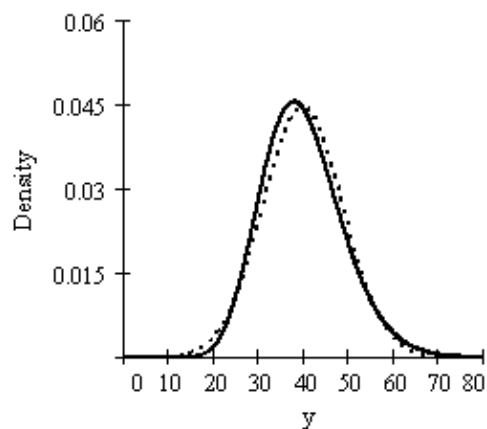
$$\chi^2(5) \approx N(5, 10)$$



$$\chi^2(10) \approx N(10, 20)$$



$$\chi^2(20) \approx N(20, 40)$$

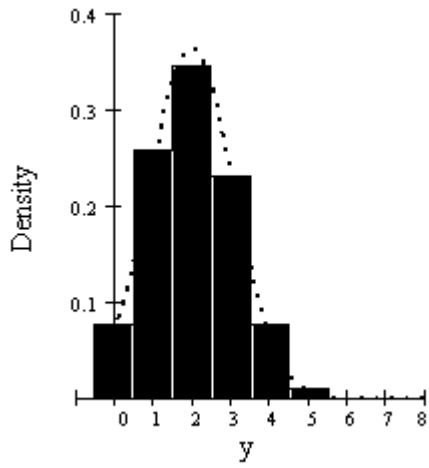


$$\chi^2(40) \approx N(40, 80)$$

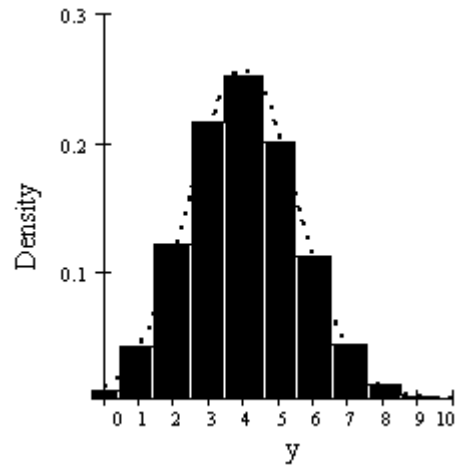
Also, the probability density function for the Binomial distribution  $B(n, p)$  can be approximated with the a normal density function with mean  $np$  and variance  $np(1-p)$ , so

$$B(n, p) \approx N(np, np(1-p)).$$

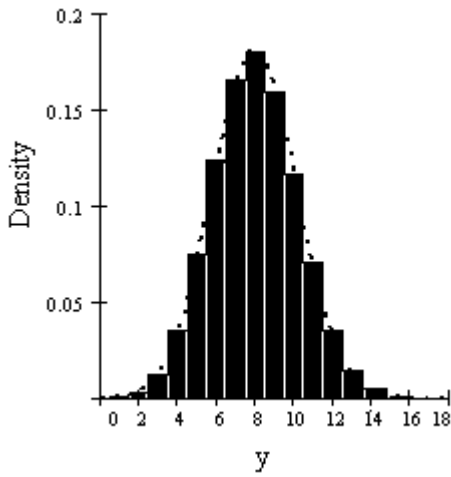
In the second set of graphs, the binomial density function is in bold and the normal probability density function is dashed.



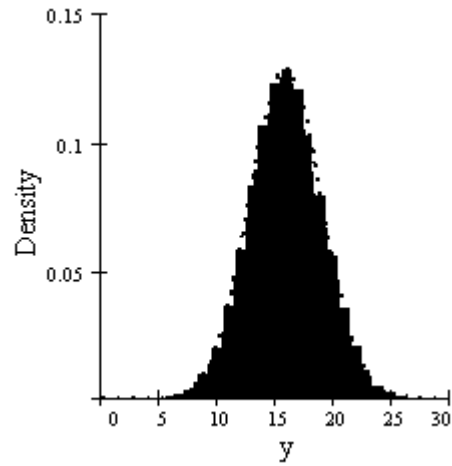
$$B(5, 0.4) \approx N(2, (5)(0.4)(0.6))$$



$$B(10, 0.4) \approx N(4, (10)(0.4)(0.6))$$



$$B(20, 0.4) \approx N(8, (20)(0.4)(0.6))$$



$$B(40, 0.4) \approx N(16, (40)(0.4)(0.6))$$