Introduction to The Central Limit Theorem

There are a number of important theorems that govern the sampling distribution of $\bar{Y}$. Principal among them stands the Central Limit Theorem. A typical presentation of the theorem is given on page 249 in *Statistics, The Exploration and Analysis of Data, 3rd*, by Devore and Peck (1997), who state it this way:

> Let $\bar{Y}$ denote the mean of the observations in a random sample of size n from a population having a mean m and standard deviation s. Denote the mean of the $\bar{Y}$ distribution by $m_{\bar{Y}}$ and the standard deviation of the $\bar{Y}$ distribution by $s_{\bar{Y}}$. Then the following rules hold:
>
> **Rule 1.** $m_{\bar{Y}} = m$
>
> **Rule 2.** $s_{\bar{Y}} = \dfrac{s}{\sqrt{n}}$ This rule is approximately correct as long as no more than 5% of the population is included in the sample.
>
> **Rule 3.** When the population distribution is normal, the sampling distribution of $\bar{Y}$ is also normal for any sample size *n*.
>
> **Rule 4. (Central Limit Theorem)**
> When *n* is sufficiently large, the sampling distribution of $\bar{Y}$ is well approximated by a normal curve, even when the population distribution is not itself normal.

The key point about the Central Limit Theorem is that it is a theorem about <u>shape</u>. The derivation for the mean and standard deviation of the sampling distribution of sample means (Rules 1 and 2) does not require an assumption of normality. Let us suppose that $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed with mean $= m$ and finite variance s^2. We now prove these two theorems about the mean and variance of the sample mean.

Theorem 1a: $E(\bar{Y}) = m$

Proof:

$$\begin{aligned}
E(\bar{Y}) &= E\left[\frac{1}{n}(Y_1 + Y_2 + Y_3 + \ldots + Y_n)\right] \\
&= \frac{1}{n}E[Y_1 + Y_2 + Y_3 + \ldots + Y_n] \\
&= \frac{1}{n}[E(Y_1) + E(Y_2) + E(Y_3) + \ldots + E(Y_n)] \\
&= \frac{1}{n}[m + m + \ldots + m]
\end{aligned}$$

$$= \frac{1}{n}\left[nm\right] = m$$

Note that independence of the Y_i's is not needed for this result.

Theorem 1b: $V\left(\bar{Y}\right) = \frac{s^{2}}{n}$

Proof:

$$V\left(\bar{Y}\right) = V\left[\frac{1}{n}\left(Y_1 + Y_2 + Y_3 + \ldots + Y_n\right)\right]$$

$$= \left(\frac{1}{n}\right)^2 V\left[Y_1 + Y_2 + Y_3 + \ldots + Y_n\right]$$

$$= \left(\frac{1}{n}\right)^2 \left[V\left(Y_1\right) + V\left(Y_2\right) + V\left(Y_3\right) + \mathrm{L} \quad + V\left(Y_n\right)\right]$$

$$= \left(\frac{1}{n}\right)^2 \left[s_1^{\ 2} + s_2^{\ 2} + s_3^{\ 2} + \mathrm{L} \quad + s_n^{\ 2}\right]$$

$$= \left(\frac{1}{n}\right)^2 \left[ns^{2}\right] = \frac{s^{2}}{n}$$

Note that $E\left(Y_i\right) = E\left(Y_j\right)$ is not required for this result. In fact, independence is not necessary, just that the *Y*'s are uncorrelated.

Before proving the Central Limit Theorem, we need an essential theorem from probability theory. This theorem will be stated but not proven.

Theorem 2: Let Y_n and *Y* be random variables with moment generating functions $m_n\left(t\right)$ and $m\left(t\right)$, respectively. If

$$\lim_{n\to\infty} m_n\left(t\right) = m\left(t\right)$$

for all real t, then the distribution function of Y_n converges to the distribution function of *Y* as $n \to \infty$.

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 $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed with mean m and finite variance s^{2}. Define the random variable U_n as follows:

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

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

$$Z_i = \frac{Y_i - m}{s} .$$

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

$$m_Z\left(t\right) = 1 + \frac{t^2}{2} + \frac{t^3}{3!}E\left(Z_i^{\,3}\right) + \cdots .$$

Also, we know that

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

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,

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

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

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

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

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

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

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

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

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

and

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

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.

Sampling Distribution of Sums

We can recast the central limit theorem as a theorem about sums of random variables also. Let us suppose that $Y_1,Y_2,\ldots,Y_n$ are independent and distributed with mean m and finite variance s^2. Let $Y^*=Y_1+Y_2+\ldots+Y_n$. Then…

$$\begin{aligned}E\left(Y^*\right)&=E\left(Y_1+Y_2+\ldots+Y_n\right)\\&=E\left(n\bar{Y}\right)\\&=nE\left(\bar{Y}\right)\\&=nm\end{aligned}$$

$$\begin{aligned}V\left(Y^*\right)&=V\left(Y_1+Y_2+\ldots+Y_n\right)\\&=V\left(n\bar{Y}\right)\\&=n^2V\left(\bar{Y}\right)\\&=n^2\left(\frac{s^2}{n}\right)=ns^2\end{aligned}$$

Then,

$$\begin{aligned}\frac{\bar{Y}-m}{\left(s/\sqrt{n}\right)}&=\frac{\bar{Y}-m}{\left(s/\sqrt{n}\right)}\cdot\frac{n}{n}=\frac{n\bar{Y}-nm}{\sqrt{n}s}\\&=\frac{Y^*-nm}{\sqrt{n}s}\end{aligned}$$

and thus Y^* must also be approximately normally distributed.

Bernoulli Trials

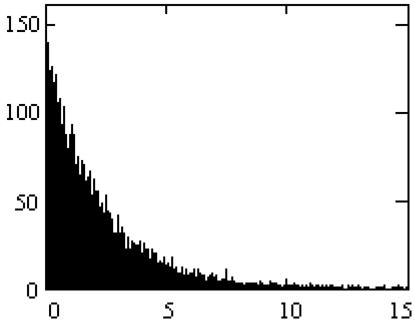
In the case that $Y_1, Y_2, \ldots, Y_n$ are distributed as counts from Bernoulli trials with probability of success *p*, $E(y_i) = p$, and $V(y_i) = p(1-p)$, it follows that

$$E\left(Y^*\right) = np \text{ and } V\left(Y*\right) = np\left(1-p\right)$$

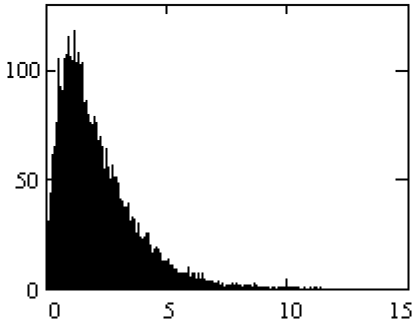
and $\dfrac{Y^* - np}{\sqrt{np(1-p)}}$ is approximately distributed as *N*(0, 1) as *n* increases without bound.

Simulations

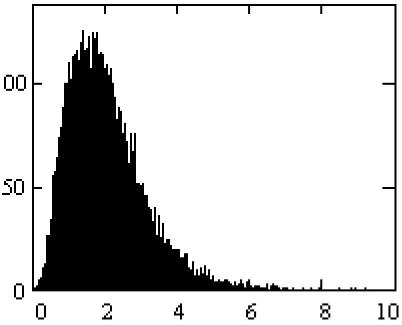
One way to "see" the Central Limit Theorem in action is through simulations. We can select *n* independent values from a given distribution, find the mean, and repeat the process several thousand times. Gathering all of the computed sample means, we can find the mean and standard deviation of these sample means and present the distribution in a histogram. A few examples are shown below. The population is $c^2\left(2\right)$ which has $m = 2$ and $s^2 = 4$. 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 $s = \dfrac{2}{\sqrt{n}}$, respectively.



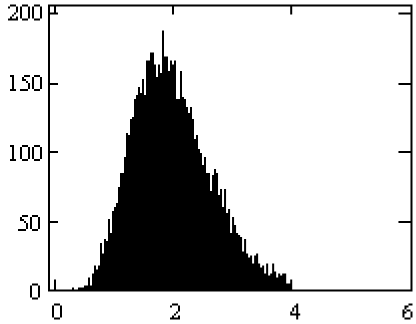
$n = 1, \bar{x} = 1.995, s = 2.007$



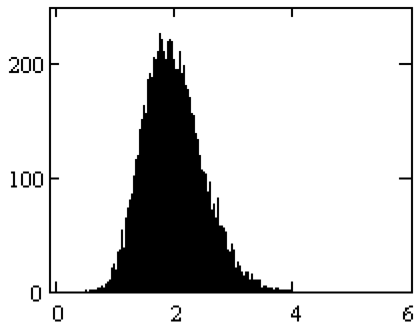
$n = 2, \bar{x} = 2.017, s = 1.434$



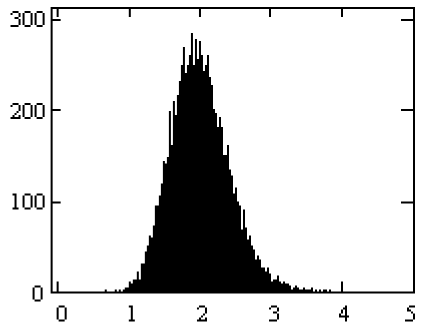
$n = 4, \bar{x} = 2.001, s = 1.005$



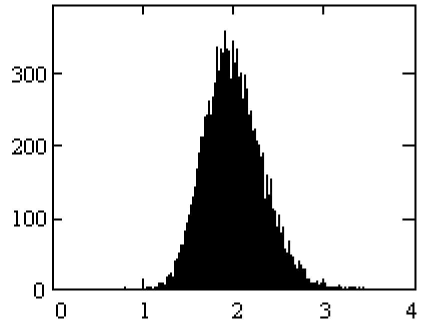
$n = 9, \bar{x} = 1.995, s = 0.664$



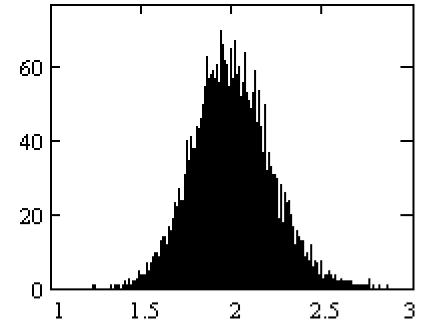
$n=16,\ \bar{x}=1.993,\ s=0.469$



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

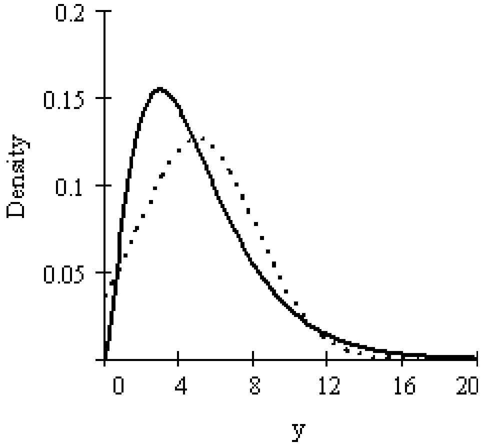


$n=40,\ \bar{x}=2.001,\ s=0.315$

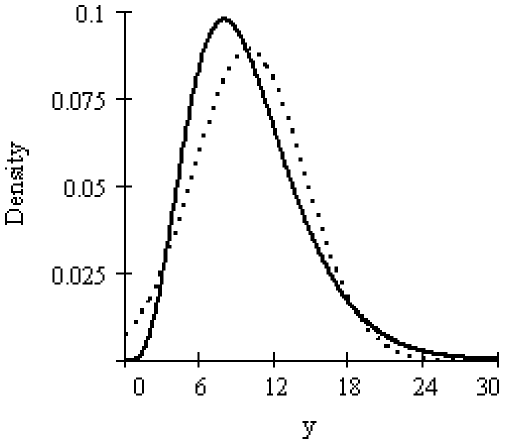


$n=100,\ \bar{x}=2.001,\ s=0.200$

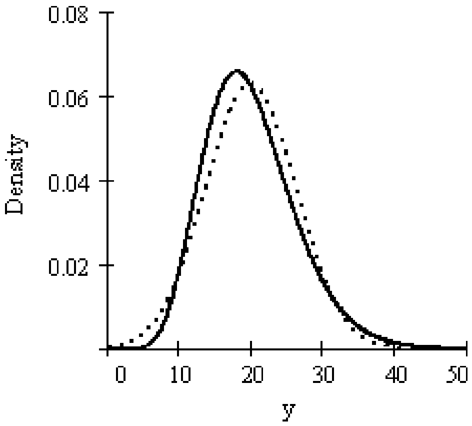
Another approach to visualizing the Central Limit Theorem is to compare the probability density functions. We expect that the distribution function for the chi-square distribution with u degrees of freedom will approach the normal distribution function with mean u and variance $2u$, that is, $c^2(u) \approx N(u,2u)$ and u 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.



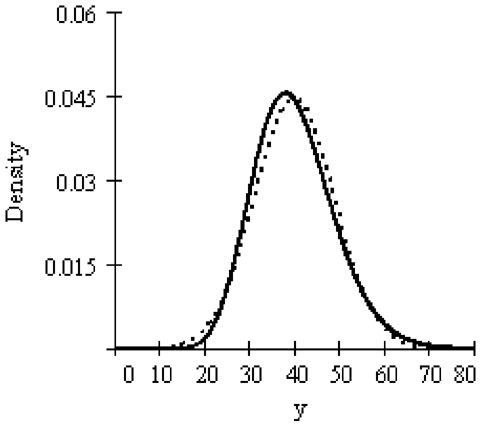
$c^2(5) \approx N(5,10)$



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



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

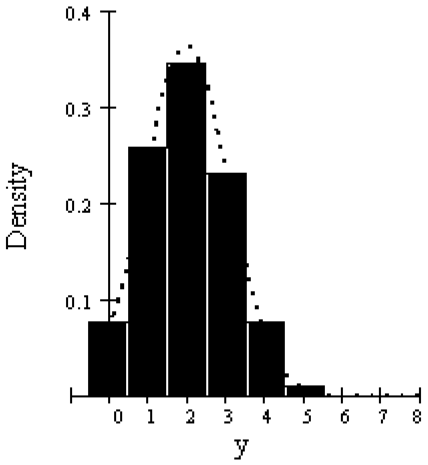


$c^2(40) \approx N(40,80)$

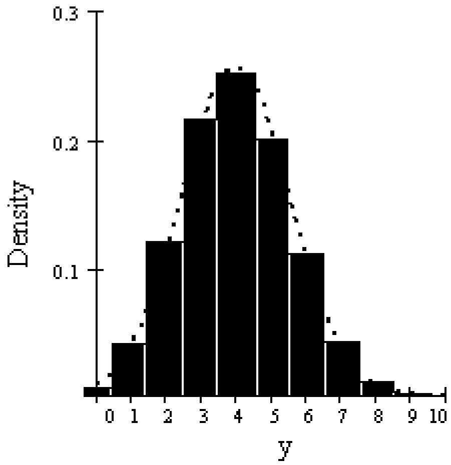
Also, the probability density function for the Binomial distribution $B(n,p)$ can be approximated with the a normal density function with mean *np* and variance $np(1-p)$, so

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

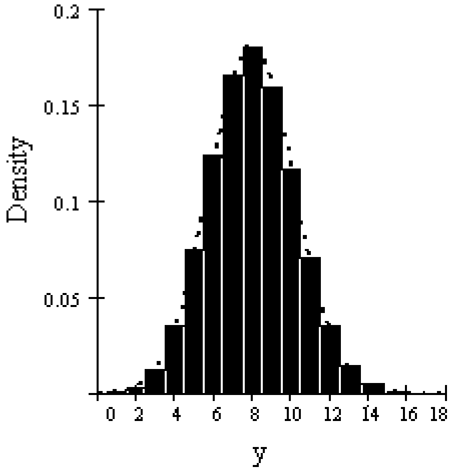
In the second set of graphs, the binomial density function is in bold and the normal probability density function is dashed.



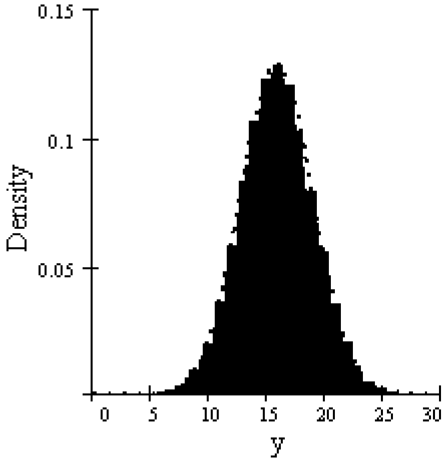
$B(5,0.4) \approx N\left(2,(5)(0.4)(0.6)\right)$



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



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



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

Confidence intervals for m.

A confidence interval for a target parameter, q, is an interval, (*L*, *U*), with endpoints *L* and *U* calculated from the sample. Ideally the resulting interval will have two properties: it will contain the population parameter a large proportion of the time, and it will be narrow. These upper and lower limits of the confidence interval are called *upper* and *lower confidence limits*. The probability that a confidence interval will contain q is called the *confidence coefficient*.

In the case of a confidence interval for the population mean, we may begin with the probability distribution for the sample mean. As we have shown above, the distribution of the sample mean is approximately normal for large *n*. That is,

$$U_n = \frac{\left(\bar{Y} - m\right)}{\left(\frac{s}{\sqrt{n}}\right)} \text{ is approximately normally distributed for large } n.$$

That being the case, for a confidence coefficient of $1 - a$ the following is true approximately for large *n*:

$$p\left(-Z < \frac{\bar{Y} - m}{\left(\frac{s}{\sqrt{n}}\right)} < Z\right) = 1 - a$$

To keep the notation simple we will construct a 95% confidence interval for the population mean. We know from our standard normal calculations that:

$$p\left(-1.96 < \frac{\overline{Y} - m}{\left(s \middle/ \sqrt{n} \right)} < 1.96\right) = 0.95$$

With a little algebra, we have:

$$p\left(-1.96 s_{\overline{Y}} < \overline{Y} - m < 1.96 s_{\overline{Y}}\right) = 0.95$$

$$p\left(1.96 s_{\overline{Y}} > m - \overline{Y} > -1.96 s_{\overline{Y}}\right) = 0.95$$

$$p\left(-1.96 s_{\overline{Y}} < m - \overline{Y} < 1.96 s_{\overline{Y}}\right) = 0.95$$

$$p\left(\overline{Y} - 1.96 s_{\overline{Y}} < m < \overline{Y} + 1.96 s_{\overline{Y}}\right) = 0.95$$

Since we typically do not know s^2, we can't evaluate $s_{\overline{Y}} = \frac{s}{\sqrt{n}}$. Instead, we substitute our estimate based on the sample to actually construct the interval. And, because we are estimating the population variance, we must use the *t*-distribution for the shape of the sampling distribution of the sample mean.

$$p\left(\overline{Y} - t^* \frac{s}{\sqrt{n}} < m < \overline{Y} + t^* \frac{s}{\sqrt{n}}\right) = 0.95$$

A simple, correct probability statement about the confidence interval is:

> the probability is 0.95 that the *process* of constructing the interval will result in an interval containing the population mean.

We usually state that we are *confident* that the interval we constructed contains the population mean.