5 Joint Probability Distributions and Random Samples

5.4 The Distribution of the Sample Mean

See Towards Central Limit Theorem

The Central Limit Theorem

The Central Limit Theorem

When the X_i's are normally distributed, so is $\overline{X}$ for every sample size n.

Even when the population distribution is highly nonnormal, averaging produces a distribution more bell-shaped than the one being sampled.

A reasonable conjecture is that if n is large, a suitable normal curve will approximate the actual distribution of $\overline{X}$. The formal statement of this result is the most important theorem of probability.

5

The Central Limit Theorem

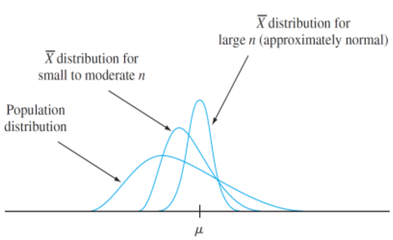
Theorem
The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma^2_{\overline{X}} = \sigma^2/n$, and T_o also has approximately a normal distribution with $\mu_{T_o} = n\mu$, $\sigma^2_{T_o} = n\sigma^2$. The larger the value of n, the better the approximation.

6

The Central Limit Theorem

Figure 5.15 illustrates the Central Limit Theorem.



The Central Limit Theorem illustrated

Figure 5.15

7

The Central Limit Theorem

According to the CLT, when n is large and we wish to calculate a probability such as $P(a \le \overline{X} \le b)$, we need only “pretend” that $\overline{X}$ is normal, standardize it, and use the normal table.

The resulting answer will be approximately correct. The exact answer could be obtained only by first finding the distribution of $\overline{X}$, so the CLT provides a truly impressive shortcut.

8

Example 26

The amount of a particular impurity in a batch of a certain chemical product is a random variable with mean value 4.0 g and standard deviation 1.5 g.

If 50 batches are independently prepared, what is the (approximate) probability that the sample average amount of impurity $\overline{X}$ is between 3.5 and 3.8 g?

According to the rule of thumb to be stated shortly, *n* = 50 is large enough for the CLT to be applicable.

9

Example 26

cont'd

$\overline{X}$ then has approximately a normal distribution with mean value $\mu_{\overline{X}} = 4.0$ and $\sigma_{\overline{X}} = 1.5/\sqrt{50} = .2121$,

so

$$P(3.5 \le \overline{X} \le 3.8) \approx P\left(\frac{3.5 - 4.0}{.2121} \le Z \le \frac{3.8 - 4.0}{.2121}\right)$$

$$= \Phi(-.94) - \Phi(-2.36)$$

$$= .1645$$

10

The Central Limit Theorem

The CLT provides insight into why many random variables have probability distributions that are approximately normal.

For example, the measurement error in a scientific experiment can be thought of as the sum of a number of underlying perturbations and errors of small magnitude.

A practical difficulty in applying the CLT is in knowing when *n* is sufficiently large. The problem is that the accuracy of the approximation for a particular *n* depends on the shape of the original underlying distribution being sampled.

11

The Central Limit Theorem

If the underlying distribution is close to a normal density curve, then the approximation will be good even for a small *n*, whereas if it is far from being normal, then a large *n* will be required.

Rule of Thumb
If *n* > 30, the Central Limit Theorem can be used.

There are population distributions for which even an *n* of 40 or 50 does not suffice, but such distributions are rarely encountered in practice.

12

The Central Limit Theorem

On the other hand, the rule of thumb is often conservative; for many population distributions, an n much less than 30 would suffice.

For example, in the case of a uniform population distribution, the CLT gives a good approximation for $n \geq 12$.

Other Applications of the Central Limit Theorem

Other Applications of the Central Limit Theorem

The CLT can be used to justify the normal approximation to the binomial distribution discussed earlier.

We know that a binomial variable X is the number of successes in a binomial experiment consisting of n independent success/failure trials with $p = P(S)$ for any particular trial. Define a new rv X_1 by

$$X_1 = \begin{cases} 1 & \text{if the 1st trial results in a success} \\ 0 & \text{if the 1st trial results in a failure} \end{cases}$$

and define $X_2, X_3, \ldots, X_n$ analogously for the other $n - 1$ trials. Each X_i indicates whether or not there is a success on the corresponding trial.

Other Applications of the Central Limit Theorem

Because the trials are independent and $P(S)$ is constant from trial to trial, the X_i's are iid (a random sample from a Bernoulli distribution).

The CLT then implies that if n is sufficiently large, both the sum and the average of the X_i's have approximately normal distributions.

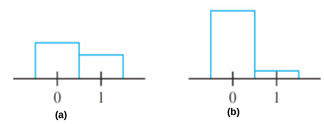
Other Applications of the Central Limit Theorem

When the X_i's are summed, a 1 is added for every S that occurs and a 0 for every F, so $X_1 + \ldots + X_n = X$. The sample mean of the X_i's is X/n, the sample proportion of successes.

That is, both X and X/n are approximately normal when n is large.

Other Applications of the Central Limit Theorem

The necessary sample size for this approximation depends on the value of p: When p is close to .5, the distribution of each X_i is reasonably symmetric (see Figure 5.19), whereas the distribution is quite skewed when p is near 0 or 1. Using the approximation only if both $np \geq 10$ and $n(1 \geq p) \geq 10$ ensures that n is large enough to overcome any skewness in the underlying Bernoulli distribution.



Two Bernoulli distributions: (a) p = .4 (reasonably symmetric); (b) p = .1 (very skewed)

Figure 5.19

Other Applications of the Central Limit Theorem

We know that X has a lognormal distribution if ln(X) has a normal distribution.

Proposition

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution for which only positive values are possible [$P(X_i > 0) = 1$]. Then if n is sufficiently large, the product $Y = X_1X_2 \ldots\ldots X_n$ has approximately a lognormal distribution.

Other Applications of the Central Limit Theorem

To verify this, note that

$$\ln(Y) = \ln(X_1) + \ln(X_2) + \cdots + \ln(X_n)$$

Since ln(Y) is a sum of independent and identically distributed rv's [the ln(X_i)s], it is approximately normal when n is large, so Y itself has approximately a lognormal distribution.

Other Applications of the Central Limit Theorem

As an example of the applicability of this result, Bury (*Statistical Models in Applied Science,* Wiley, p. 590) argues that the damage process in plastic flow and crack propagation is a multiplicative process, so that variables such as percentage elongation and rupture strength have approximately lognormal distributions.

21