13.0 Central Limit Theorem

- Discuss Midterm/Answer Questions
- Box Models
- Expected Value and Standard Error
- Central Limit Theorem

13.1 Box Models

A Box Model describes a process in terms of making repeated draws, with replacement, from a box containing numbers.

Since draws are made with replacement, the outcomes in a series of draws are independent. The value on the first draw does not affect the value on the second.

Box models describe: repeated rolls of a die, repeated tosses of a coin (fair or unfair), drawing a random sample of people and getting their heights or whom they plan to vote for.

For a box model, the expected value is the average of the numbers in the box.

For categorical data, such as H or T in coin-tossing, or voting for Bush or Kerry, one averages zeroes and ones. When averaging zeroes and ones, the result is just a proportion, or the total number of people voting for Kerry divided by the total number of people whom you ask.

The Law of Averages says that if one makes many draws from the box and averages the results, that average will converge to the expected value of the box.

But the Law of Averages says nothing about the outcome on the next draw.

13.2 Expected Value and Standard Error

Suppose that a box contains B numbers, $X_1, \ldots, X_B$. Then the expected value of the box is is

$$EV = \bar{X} = \frac{1}{B}\sum_{i=1}^{B} X_i$$

and the standard deviation of the box is

$$\begin{aligned} sd &= \sqrt{\frac{1}{n}[(X_1 - EV)^2 + \cdots + (X_B - EV)^2]} \\ &= \sqrt{\left(\frac{1}{n}\sum_{i=1}^{B} X_i^2\right) - EV^2}. \end{aligned}$$

This is just the same as calculating the mean and standard deviation for a list of numbers.

The standard error for the average of n draws from a box (with replacement) is:

$$se = \frac{sd}{\sqrt{n}}.$$

The standard error is the likely size of the difference between the average of n draws from the box and the expected value of the box.

Note that as $n \to \infty$, the standard error se goes to zero. This is a formal statement of the Law of Averages. It means that the sample average is a good estimate of the average in the box, and the accuracy of the estimate improves as you take more and more draws from the box.

The standard error for the sum of n draws from a box (with replacement) is:

$$se = \sqrt{n} \times sd.$$

Analgously, the standard error is the likely size of the difference between the sum of n draws from the box and n times the expected value of the box.

Note that as $n \to \infty$, this standard error does not go to zero. This means that as the number of draws increases, the likely difference between the sum of the draws and nEV gets larger, rather than smaller.

This concern about the sum, rather than the average, arises in contexts such as investment or gambling, where the total return from multiple trials is important, not the average return.

13.3 The Central Limit Theorem

The Central Limit Theorem is one of the high-water marks of mathematical thinking. It was worked upon by James Bernoulli, Abraham de Moivre, and Alan Turing. Over the centuries, the theory improved from special cases to a very general rule.

Essentially, the Central Limit Theorem allows one to describe how accurately the Law of Averages works. Most people have a good intuitive understanding of the Law of Averages, but in many cases it is important to determine whether a particular size of deviation between the sample mean and the (usually unknown) expected value is probable or improbable. That is, what is the chance that the sample average is more than d away from the true EV?

Formally, the Central Limit Theorem for averages says:

$$\frac{\bar{X} - EV}{sd/\sqrt{n}} \mathrel{\dot{\sim}} N(0, 1)$$

where $\bar{X}$ is the average of n draws, EV is the expected value of the box, and sd is the standard deviation of the box.

This means that the left-hand side is a random number that is approximately normal with mean zero and standard deviation one.

The approximation gets better as n gets larger.

Modifications of this formula hold for many other situations, e.g., when there is a little dependence, or when the box changes from draw to draw.

A version of the Central Limit holds for sums:

$$\frac{n\bar{X} - nEV}{\sqrt{n}\, sd} \mathrel{\dot{\sim}} N(0,1).$$

Note that $n\bar{X}$ is just the sum of the draws from the box. (This should be obvious to everyone.)

This formula is useful when calculating the chance of winning a given amount of money when gambling, or getting more than a specific score on a test.

With these two central limit formulas, one can answer all sorts of practical questions.

Problem 1: You want to estimate the average income of people in Durham. You take a random sample of 100 households, and find that $\bar{X}$ is \$42,000 and the sample *sd* is \$5,000. What is the (approximate) probability that the true mean household income in Durham is more than \$42,500?

- What is the box model for this problem?
- What is the expected value?
- What is the standard deviation?

Note that in order to solve this, we have to assume that the standard deviation of the sample is equal to the *sd* of the box. In practice, there is a very easy way to handle this, but we will not talk about that until later in the course.

$$
\begin{aligned}
\mathbf{P}[EV > 42,500] &= \mathbf{P}[-EV < -42,500] \\
&= \mathbf{P}[\bar{X} - EV < \bar{X} - 42,500] \\
&= \mathbf{P}[\frac{\bar{X} - EV}{sd/\sqrt{n}} < \frac{\bar{X} - 42,500}{sd/\sqrt{n}}] \\
&\doteq \mathbf{P}[Z < \frac{\bar{X} - 42,500}{sd/\sqrt{n}}] \\
&= \mathbf{P}[Z < (42,000 - 42,500)/(5000/10)] \\
&= \mathbf{P}[Z < -1]
\end{aligned}
$$

From the standard normal table, we know this has chance (1/2) (100 - 68.27) = 15.865%, so the probability of the estimate being too low by \$500 is just .15865.

Problem 2: You are playing Red and Black in roulette. (A roulette wheel has 38 pockets; 18 are red, 18 are black, and 2 are green—the house takes all the money on green). You pick either red or black; if the ball lands in the color you pick, you win a dollar. Otherwise you lose a dollar.

Suppose you make 100 plays. What is the chance that you lose $10 or more?

What is the box model?

.

There are 38 tickets, and 18 are labelled 1 and the 20 are labelled -1.

So the expected value of the box is

$$\begin{aligned} EV &= \frac{1}{38}[1+1+\cdots+1+(-1)+(-1)+\cdots+(-1)] \\ &= \frac{1}{38}[-2] \\ &= -1/19. \end{aligned}$$

The standard deviation of the box is

$$\begin{aligned} sd &= \sqrt{(\frac{1}{38}\sum i = 1^{38} X_i^2) - EV^2} \\ &= \sqrt{1-(-1/19)^2} \\ &= .998614. \end{aligned}$$

The probability of losing more than $10 or more in 100 plays is

$$\begin{aligned}
\mathbf{P}[\mathbf{sum} < -10] &= \mathbf{P}[\mathbf{sum} - nEV < -10 - nEV] \\
&= \mathbf{P}[\frac{\mathbf{sum} - nEV}{\sqrt{n}sd} < \frac{-10 - nEV}{\sqrt{n}sd}] \\
&\doteq \mathbf{P}[Z < \frac{-10 - nEV}{\sqrt{n}sd}] \\
&= \mathbf{P}[Z < [-10 - (100)(-1/19)]/(10 * .998614)] \\
&= \mathbf{P}[Z < .47434].
\end{aligned}$$

From the standard normal table, the chance of this is about 1/2(100 - 34.73)%, so the probability is .32635.