Minitab®
QUALITY. ANALYSIS. RESULTS.

Tumbling Dice & Birthdays

Understanding the Central Limit Theorem

Mark Twain famously quipped that there were three ways to avoid telling the truth: lies, damned lies, and statistics. The joke works because statistics frequently seems like a black box—it can be difficult to understand how statistical theorems make it possible to draw conclusions from data that, on their face, defy easy analysis.

But because data analysis plays a critical role in everything from jet engine reliability to determining the shows we see on television, it's important to acquire at least a basic understanding of statistics. And one of the most important concepts to understand is the central limit theorem.

In this article, we will explain the central limit theorem and show how to demonstrate it using common examples, including the roll of a die and the birthdays of Major League Baseball players.

Defining the Central Limit Theorem

A typical textbook definition of the central limit theorem goes something like this:

As the sample size increases, the sampling distribution of the mean, $\overline{X}$, can be approximated by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ where:

> μ is the population mean,
> σ is the population standard deviation,
> *n* is the sample size.

In other words, if we repeatedly take independent random samples of size *n* from any population, then when *n* is large, the distribution of the sample means will approach a normal distribution.

How large is large enough? Generally speaking, a sample size of 30 or more is considered to be large enough for the central limit theorem to take effect. The closer the population distribution is to a normal distribution, the fewer samples needed to demonstrate the theorem. Populations that are heavily skewed or have several modes may require larger sample sizes.

Why Does It Matter?

The field of statistics is based upon the fact that it is rarely feasible or practical to collect all of the data from an entire population. Instead, we can gather a subset of data from a population, and then use statistics for that sample to draw conclusions about the population.

For example, we can collect random samples from an industrial process, then use the means of our samples to make conclusions about the stability of the overall process.

Two common characteristics used to define a population are the mean and standard deviation. When data follow a normal distribution, the mean indicates where the center of that distribution is, and the standard deviation reveals the spread.

Imagine you are getting the results of a test you took. In addition to receiving your own results, you also want to know your peers' average score. However, if the test scores do not follow a normal distribution, the average could be misleading.

The central limit theorem is remarkable because it implies that, no matter what the population distribution looks like, the distribution of the sample means will approach a normal distribution. The theorem also allows us to make probability statements about the possible range of values the sample mean may take. This is because the normal distribution has a useful property called the empirical rule. The rule states that for data which follow a normal distribution:

68% of the data fall within 1σ of μ
95% of the data fall within 2σ of μ
99.7% of the data fall within 3σ of μ

Watching the Theorem Work

Seeing how it can be applied makes the central limit theorem easier to understand, and we will demonstrate the theorem using dice and also using birthdays.

Example 1: Tumbling Dice

Dice are ideal for illustrating the central limit theorem. If you roll a six-sided die, the probability of rolling a one is 1/6th, a two is 1/6th, a three is also 1/6th, etc. The probability of the die landing on any one side is equal to the probability of landing on any of the other five sides.

In a classroom situation, we can carry out this experiment using an actual die. Alternatively, we can save time by using Minitab's **Calc > Random Data > Integer** menu. To get an accurate representation of the population distribution, let's roll the die 500 times. When we use a histogram to graph the data, we see that—as expected—the distribution looks fairly flat. It's definitely not a normal distribution (Fig. 1).

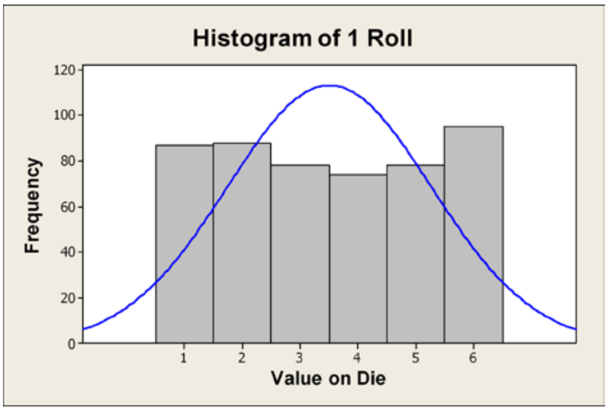
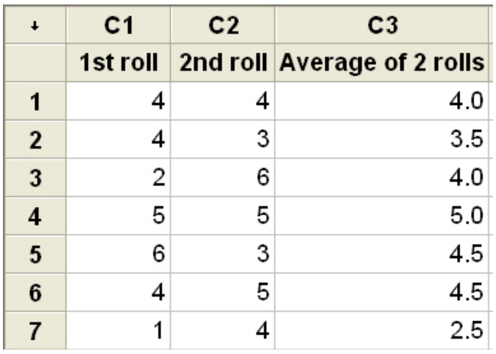


Figure 1. Because the odds of landing on all sides of a six-sided die are equal, the distribution of 500 die rolls is flat.

Let's take more samples and see what happens to the histogram *of the averages* of those samples.

This time we will roll the die twice, and repeat this process 500 times. Again we can use **Calc > Random Data > Integer** to "roll" the die for us. We can then use **Calc > Row Statistics** to compute the average of each pair (Fig. 2).



↓	C1	C2	C3
	1st roll	2nd roll	Average of 2 rolls
1	4	4	4.0
2	4	3	3.5
3	2	6	4.0
4	5	5	5.0
5	6	3	4.5
6	4	5	4.5
7	1	4	2.5

Figure 2. Minitab makes it easy to generate die-rolling data, then calculate the averages.

We can then create a histogram of these averages to view the shape of their distribution (Fig. 3). Although the blue normal curve does not accurately represent the histogram, the profile of the bars is looking more bell-shaped. Now let's roll the die five times and compute the average of the five rolls, again repeated 500 times. Then, let's repeat the process rolling the die 10 times, then 30 times.

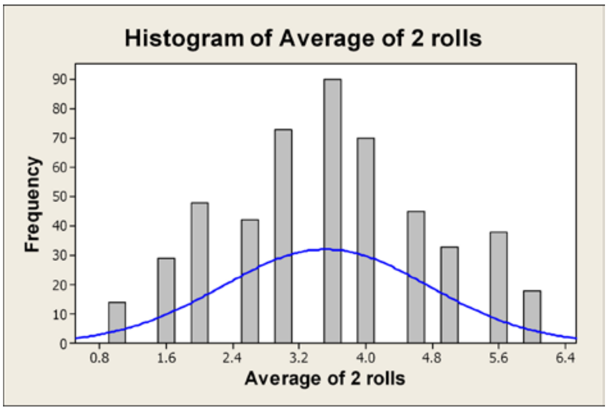


Figure 3. The distribution of 500 averages for two rolls of a die begins to resemble the familiar bell shape of a normal distribution.

The histograms for each set of averages (Fig. 4) show that as the sample size, or number of rolls, increases, the distribution of the averages comes closer to resembling a normal distribution. In addition, the variation of the sample means decreases as the sample size increases.

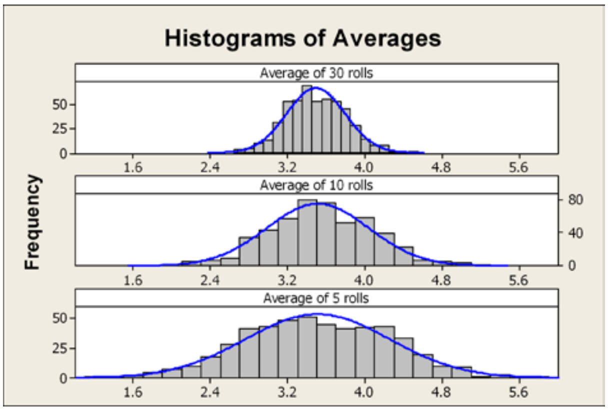


Figure 4. As the number of rolls of the die increases, the distribution of averages approaches a normal distribution.

The central limit theorem states that for a large enough *n*, $\overline{X}$ can be approximated by a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

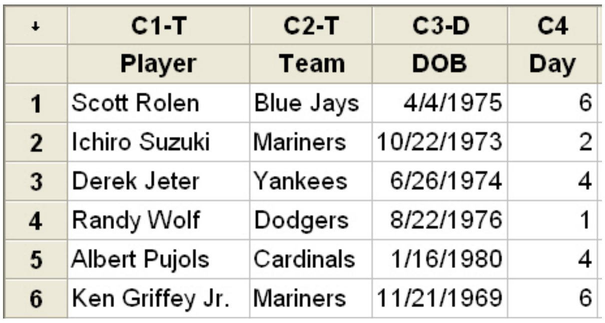
The population mean for a six-sided die is (1+2+3+4+5+6)/6 = 3.5 and the population standard deviation is 1.708. Thus, if the theorem holds true, the mean of the thirty averages should be about 3.5 with standard deviation $1.708/\sqrt{30} = 0.31$. Using the dice we "rolled" in Minitab, the average of the 30 averages, depicted in Figure 4, is 3.49 with standard deviation 0.30, which is very close to the calculated approximations.

Example 2: Birthdays

Now let's demonstrate the central limit theorem using birthdays. You'll recall that the sides of dice have an equal probability. Contrary to popular belief, there is not necessarily an equal chance of being born on a Sunday instead of a Monday or any other day of the week. Currently, the most popular day for babies to be born in the United States is Wednesday—there are 15.4% more births on Wednesday than the average day. And from 1990 to 2006, Tuesday was the most popular birth day.

To demonstrate the central limit theorem using birthdays, we first need to collect some birthdates. Students could gather the birthdays of their friends, families, and colleagues. We will use the birthdays of the more than 700 Major League Baseball players, which are available on mlb.com.

Of course, most birthday information won't include the day of the week. But using Minitab's **Data > Extract from Date/Time > To Numeric**, we can easily find out which day each baseball player was born (Fig. 5). For example, Minitab can tell us that Derek Jeter, whose birthday is June 26, 1974, was born on a Wednesday.



↓	C1-T	C2-T	C3-D	C4
	Player	Team	DOB	Day
1	Scott Rolen	Blue Jays	4/4/1975	6
2	Ichiro Suzuki	Mariners	10/22/1973	2
3	Derek Jeter	Yankees	6/26/1974	4
4	Randy Wolf	Dodgers	8/22/1976	1
5	Albert Pujols	Cardinals	1/16/1980	4
6	Ken Griffey Jr.	Mariners	11/21/1969	6

Figure 5. We can use Minitab to extract days of the week from our birthday data.

If we look at the histogram for the population of baseball players (Fig. 6), where one equals Sunday, two equals Monday, and so on, we can see that the birth days do not follow a normal distribution, and Tuesday (3) is the most popular day.

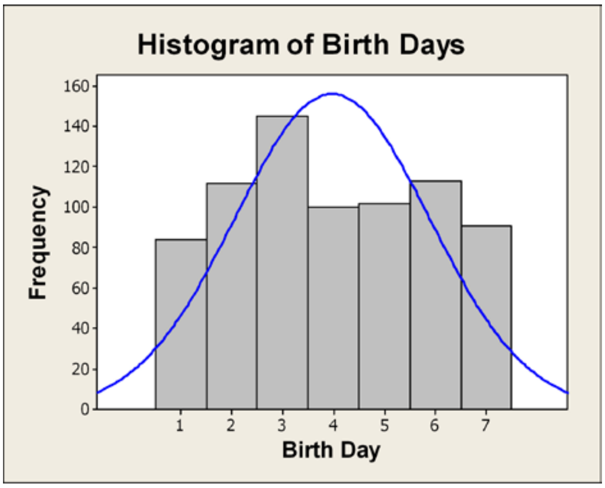


Figure 6. This histogram shows that Tuesday is the most popular birth day for Major League Baseball players.

Just like we did in the dice experiment, we will now create samples of size two, randomly sampling two players, then another two, and so forth. Let's take 100 samples total. To randomly sample players' birth days from the data in the worksheet, we can use **Calc > Random Data > Sample From Columns**.

Then let's compute the average birth day for each sample of size two using **Calc > Row Statistics**.

We will repeat the random sampling and averaging for five players, then ten players, then thirty players, and create histograms for each set of averages.

In the original histogram of more than 700 baseball players, we saw a non-normal distribution. When we look at the histograms of the averages (Fig. 7), we see they very quickly resemble a normal distribution and that the variation decreases as the sample size increases.

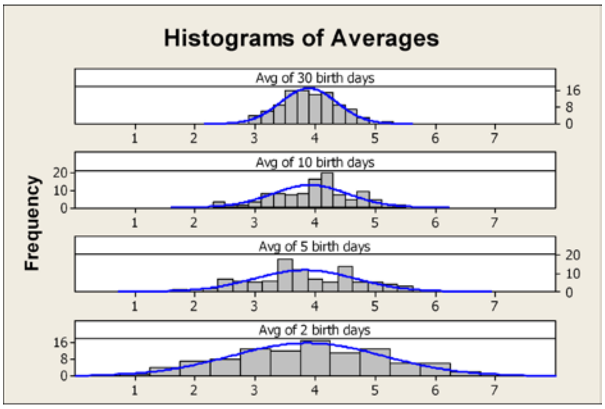


Figure 7. The central limit theorem in action: as sample size increases, the distribution of the averages more closely resembles a normal distribution.

Conclusion

The central limit theorem enables us to approximate the sampling distribution of $\overline{X}$ with a normal distribution. This idea may not be frequently discussed outside of statistical circles, but it's an important concept. And we can make it easier to understand through simple demonstrations using dice, birthdays, dates on coins, airline flight delays, or cycle times.

With an improved understanding of the central limit theorem and other statistical concepts, students with eager minds will soon find it easier to distinguish between lies, damned lies, and the truth that lies behind good statistics.

Michelle Paret
Product Marketing Manager, Minitab Inc.

Eston Martz
Senior Creative Services Specialist, Minitab Inc.

QUALITY. ANALYSIS. RESULTS.

Visit www.minitab.com for more information about statistics.

© 2009 Minitab Inc. Reprinted with permission. Minitab®, Quality Companion by Minitab®, Quality Trainer by Minitab®, Quality. Analysis. Results® and the Minitab® logo are all registered trademarks of Minitab, Inc., in the United States and other countries.