



## M5-04: Discovering The Central Limit Theorem in Python

Part of the "Polling, Confidence Intervals, and the Normal Distribution" Learning Badge

Video Walkthrough: <https://discovery.cs.illinois.edu/m5-04/>

### The Central Limit Theorem

The central limit theorem informs us that the **more times we sample** from **any distribution** the sum of those samples will tend towards a normal distribution. To discover this property ourselves, we need to both **(1)** sample from a distribution and then **(2)** aggregate the samples by summing the results together. Let's tackle each part independently!

### Part 1: Sampling from a Distribution

A distribution is the set and frequency of all possible outcomes. One simple distribution is that of a game of roulette where you always bet on red. Knowing a roulette table has 18 red spaces, 18 black spaces, and 2 green spaces, let's create our roulette distribution:

Distribution of Winnings in Roulette Betting on Red		
Outcome	Frequency	Value
Red		
Not-Red		

Note that this distribution has only two results, so it is a textbook **Bernoulli Distribution** with  $p=$ \_\_\_\_\_. We can use Python to sample from this distribution:

<b>Python:</b>	
<b>Description:</b>	Sample from a Bernoulli Distribution with $p=$ _____.

**Insight:** We could also have simulated this distribution, by simulating drawing for a queen many times with a **drawForQueen** function:

```
1 def playRoulette(n):
2     gamesWon = 0
3     for i in range(n):
4         result = random.randint(0, 38)
5         if result <= 17: # Assume 0-17 are red results
6             gamesWon = gamesWon + 1
7     return gamesWon / n # % of games won for `n` games played
```

However, since we have an exact mathematical model, the simulation is not needed.



## M5-04: Discovering The Central Limit Theorem in Python

Part of the “Polling, Confidence Intervals, and the Normal Distribution” Learning Badge

Video Walkthrough: <https://discovery.cs.illinois.edu/m5-04/>


### Part 2: Aggregating Samples from a Distribution

To discover the central limit theorem, we now must simulate sampling the distribution various numbers of times.

**Puzzle #1:** Write the simulation that **runs 10,000 times** that **samples the “Roulette Betting on Red” distribution only one time each simulation:**




<b>Python:</b>	
<b>Description:</b>	Simulates sampling from the “Roulette Betting on Red” one time.

When using `df.plot.hist()` to create a histogram of the distribution, what is the resultant histogram?

<b>Histogram with 1 sample from the distribution (k=1):</b>	
---	--

**Puzzle #2:** Modify Puzzle #1 to find the sum of **k** samples from the distribution, where **k** is a number we can configure (ex:  $k=10$ ,  $k=100$ ,  $k=10000$ , etc).

**Puzzle #3:** Find the histograms for increasingly large values of **k**:

 Simulation with <b>k=10</b>	 Simulation with <b>k=100</b>	 Simulation with <b>k=10000</b>
--	---	---

...what trends do you see as **k** gets larger?