Week Five: The Central Limit Theorem

CS 217

Estimation

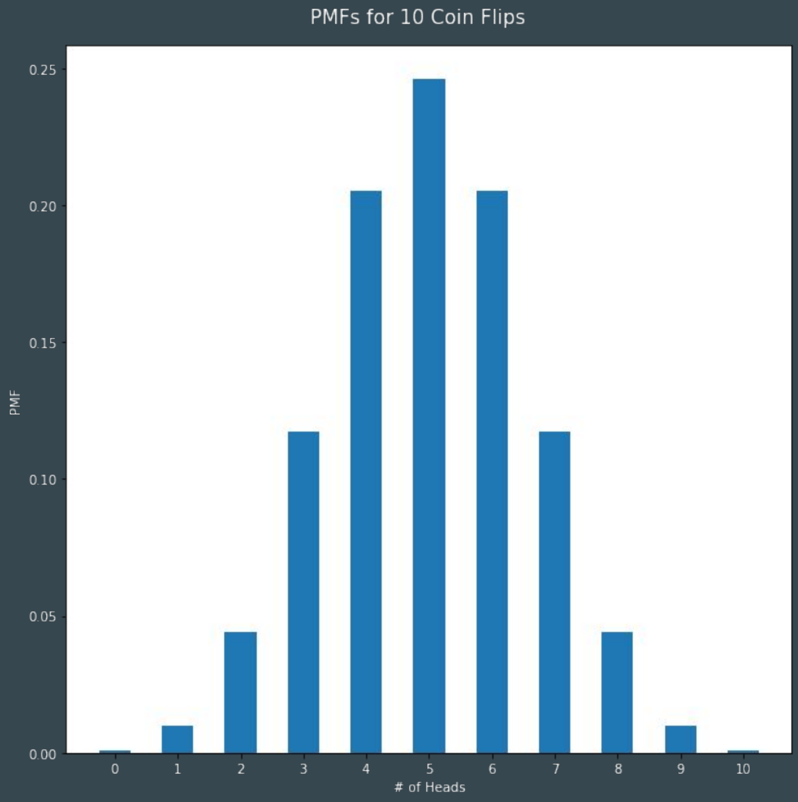
- Thus far we have been dealing with distributions where we know the inherent characteristics, such as the mean.
- For example we inherently know that a coin flip is a Bernoulli trial with a probability of 0.5.
- We also know that **randomness** is inherent in probability
- If I flip ten coins, what is the expected number of times I will land on heads?

Estimation

- Thus far we have been dealing with distributions where we know the inherent characteristics, such as the mean.
- For example we inherently know that a coin flip is a Bernoulli trial with a probability of 0.5.
- We also know that **randomness** is inherent in probability
- If I flip ten coins, what is the expected number of times I will land on heads?
    - *Five*
- If I flip ten coins, what is my probability of getting five heads? How do we calculate this?

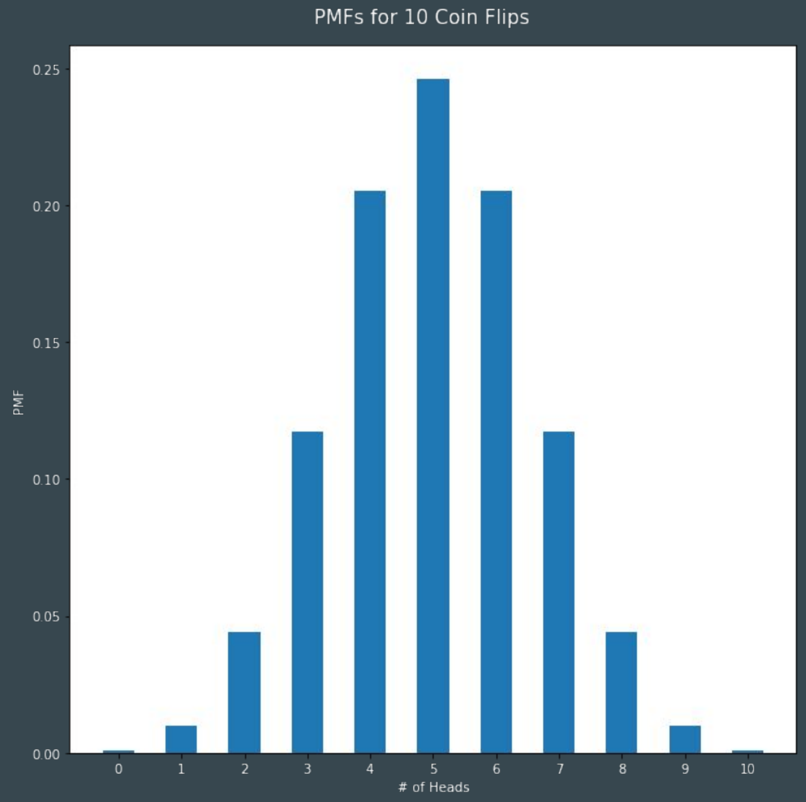
Estimation

- We can find the **probability mass function** for this.
- Here, there is only a ~25% chance that we will get our expected value of 5 flips.



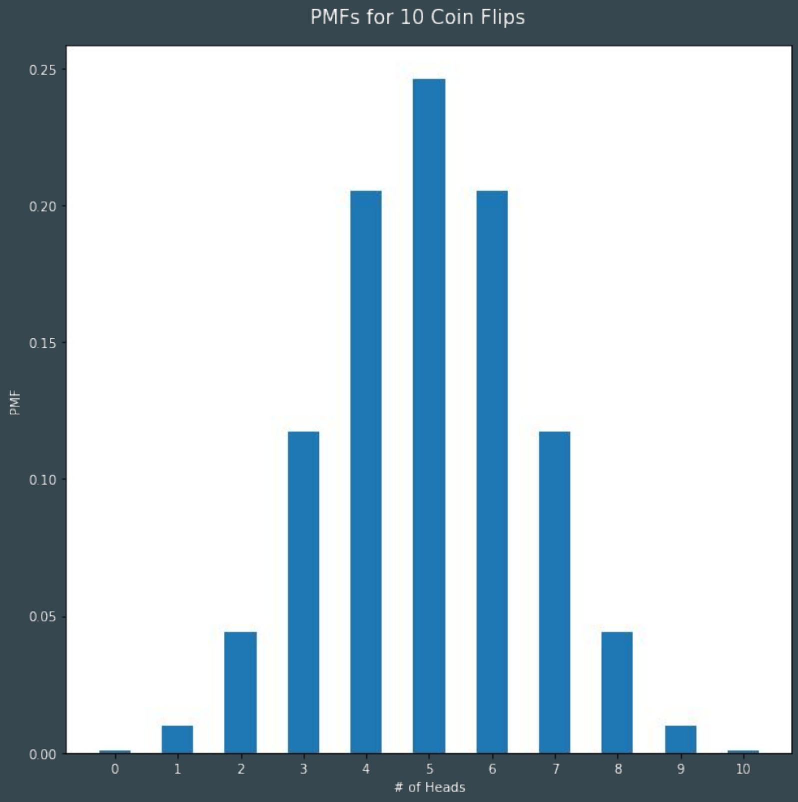
Estimation

- Let's give ourselves a little room for error here. What is the probability that we will be *close* to getting the true estimate? Say, that if we expect ten heads we will get somewhere between 4 and 6 heads?
- This means that our sample with be within 0.1 of the true probability, since our mean is 5 (10 * 0.5). The lower cutoff for our sample is 0.4 (10 * 0.4) and the upper cutoff for our sample is 0.6 (10 * 0.6)



Estimation

- The probability that we'll get between 4 and 6 heads is the sum of their respective PMFs, or around 65.6%.



Result	PMF
4	0.205
5	0.246
6	0.205

Estimation

- Say we now flip 50 coins. With a probability of 0.5, we expect that we'll land on 25 heads.
- What is the probability that we will be *close* to our estimate, and be within 0.1 of our true probability?
- What is the lower probability cutoff?
- What is the number of heads that we would 'accept' given our lower probability cutoff?
- What is the upper probability cutoff?
- What is the number of heads that we would 'accept' given our upper probability cutoff?

Estimation

- Say we now flip 50 coins. With a probability of 0.5, we expect that we'll land on 25 heads.
- What is the probability that we will be *close* to our estimate, and be within 0.1 of our true probability?
- What is the lower probability cutoff?
    - *0.4*
- What is the number of heads that we would 'accept' given our lower probability cutoff?
- What is the upper probability cutoff?
- What is the number of heads that we would 'accept' given our upper probability cutoff?

Estimation

- Say we now flip 50 coins. With a probability of 0.5, we expect that we'll land on 25 heads.
- What is the probability that we will be *close* to our estimate, and be within 0.1 of our true probability?
- What is the lower probability cutoff?
    - *0.4*
- What is the number of heads that we would 'accept' given our lower probability cutoff?
    - *50 * 0.4 = 20*
- What is the upper probability cutoff?
- What is the number of heads that we would 'accept' given our upper probability cutoff?

Estimation

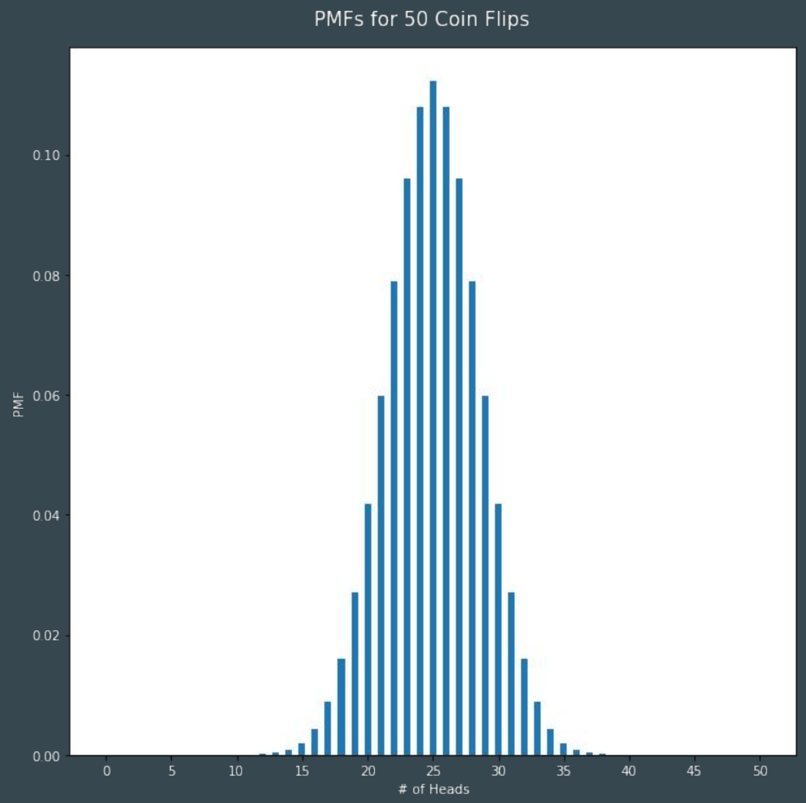
- Say we now flip 50 coins. With a probability of 0.5, we expect that we'll land on 25 heads.
- What is the probability that we will be *close* to our estimate, and be within 0.1 of our true probability?
- What is the lower probability cutoff?
  - *0.4*
- What is the number of heads that we would 'accept' given our lower probability cutoff?
  - *50 * 0.4 = 20*
- What is the upper probability cutoff?
  - *0.6*
- What is the number of heads that we would 'accept' given our upper probability cutoff?

Estimation

- Say we now flip 50 coins. With a probability of 0.5, we expect that we'll land on 25 heads.
- What is the probability that we will be *close* to our estimate, and be within 0.1 of our true probability?
- What is the lower probability cutoff?
    - *0.4*
- What is the number of heads that we would 'accept' given our lower probability cutoff?
    - *50 * 0.4 = 20*
- What is the upper probability cutoff?
    - *0.6*
- What is the number of heads that we would 'accept' given our upper probability cutoff?
    - *50 * 0.6 = 30*

Estimation

- So now, we want to see the probability that we land in between 20 and 30 heads, by finding the PMFs of each of those values.
- The probability that we get between 20 and 30 heads is 88.1%
- This compares to 65.6% for ten coin flips.
- The more samples we have, the higher chance we have of being close to the true expected value.



Law of Large Numbers

- Indeed, if we were to keep increasing the number of trials, the higher chance we would have of getting a result close to the expected value.

Number of Samples	Lower Cutoff	Upper Cutoff	Probability
10	4	6	65.6%
50	20	30	88.1%
100	40	60	96.4%
500	200	300	99.9%
1000	400	600	100%

Law of Large Numbers

- Indeed, if we were to keep increasing the number of trials, the higher chance we would have of getting a result close to the expected value.
- This works if we get more rigorous with our estimation and only allow an error of 0.01, as opposed to an error of 0.1.

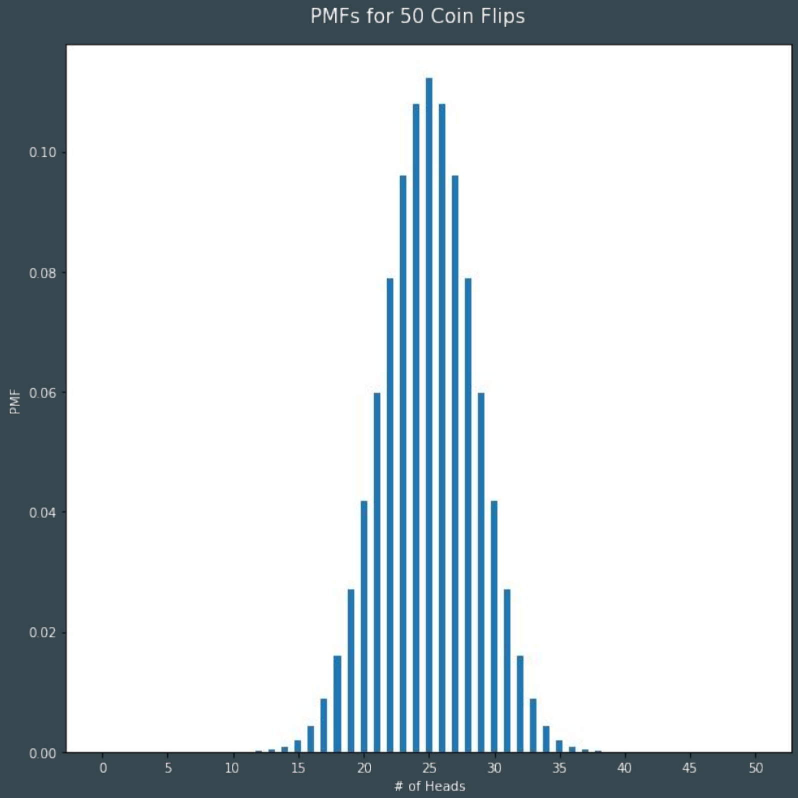
Number of Samples	Lower Cutoff	Upper Cutoff	Probability
100	49	51	23.5%
500	245	255	37.7%
1000	490	510	49.3%
5000	2450	2550	84.6%
10000	4900	5100	95.5%

Law of Large Numbers

- This phenomenon is called the **law of large numbers**.
- The **law of large numbers** states that as n, the number of trials in a simulation, goes up, the probability that our simulated mean is close to the true mean converges to 100%.
- Note that for this to be true, our trials must be **independent** and **identically distributed** (or drawing from the same distribution).

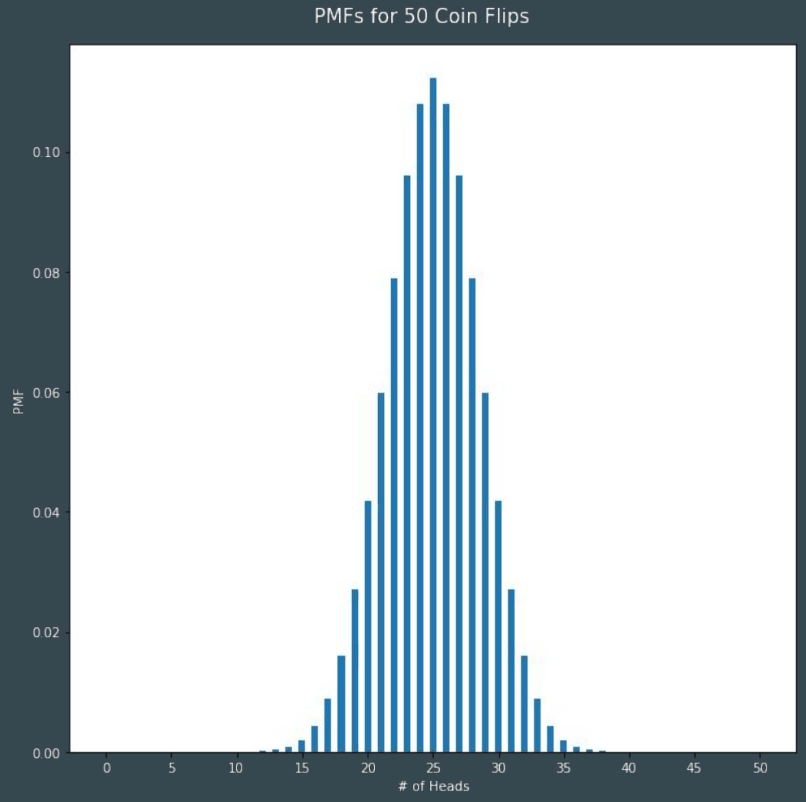
Normal Approximation

- Note that this distribution looks a lot like a normal distribution!
- Indeed, for the **binomial distribution**, we can use the normal approximation to simulate the probability of a binomial distribution.



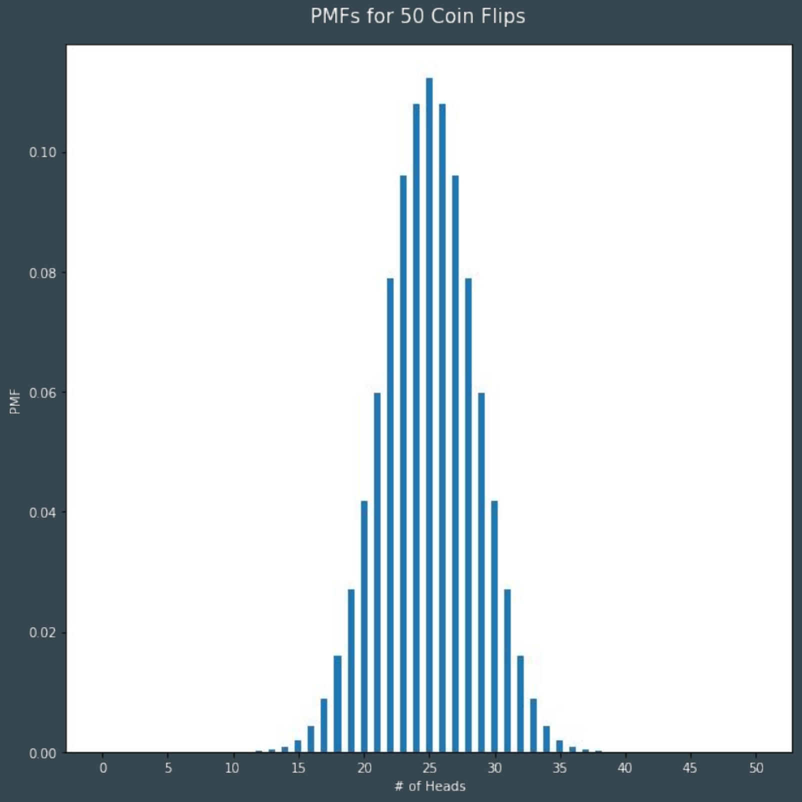
Normal Approximation

- Let's use the pictured sample here, where we have 50 coin flips.
- This is a Binomial distribution with p=0.5 and n=50 (note that we were referencing 50 trials from a Bernoulli distribution earlier - this is the same thing).
- This Binomial distribution would have a variance of n * p * 1 - p, or 50 * 0.5 * 0.5, or 12.5
- It would thus have a standard deviation of ~3.53



Normal Approximation

- Say we wanted to find the probability of getting 23 or less coin flips.
- We could find the CDF from the binomial distribution at 23 (or add up the PMFs from 0 - 23).
- This value is about 0.336.



Normal Approximation

- Or we could model a normal distribution with a mean of 25 and a standard deviation of 3.53.
- We could then find the CDF of 23 in this normal distribution
- We do this by applying a **continuity correction** and actually finding the CDF of 23.5 in this distribution.
- We will prove this out via code, but the CDF is remarkably close with this method - also 3.536.
- The normal approximation applies when n * p >= 5, and n*(1-p) >= 5

Normal Approximation

- The normal approximation was especially useful before we had computing power to easily get the PMF and CDF of a binomial distribution.
- It is much easier to do a Z-score calculation by hand then it is to the permutations and combinations required to find the PMF values of a binomial distribution by hand.
- This isn't as relevant anymore, but it's still an important application (and proof) of the **Central Limit Theorem**, which we'll learn about shortly.

Estimation

- Let's say that I am cold-calling new customers about my new banking product.
- Inherently my probability of getting a given customer to give me their email address so I can send them more information about the product is 20%.
- This is simply a Bernoulli trial with a 20% chance of success.
- Here, the expected probability of success is 20%
- The variance for a Bernoulli distribution is p * (1 - p). In our case that is 0.2 * 0.8, or 0.16.
- The standard deviation for a Bernoulli distribution is the square root of the variance, which in this case is 0.4.

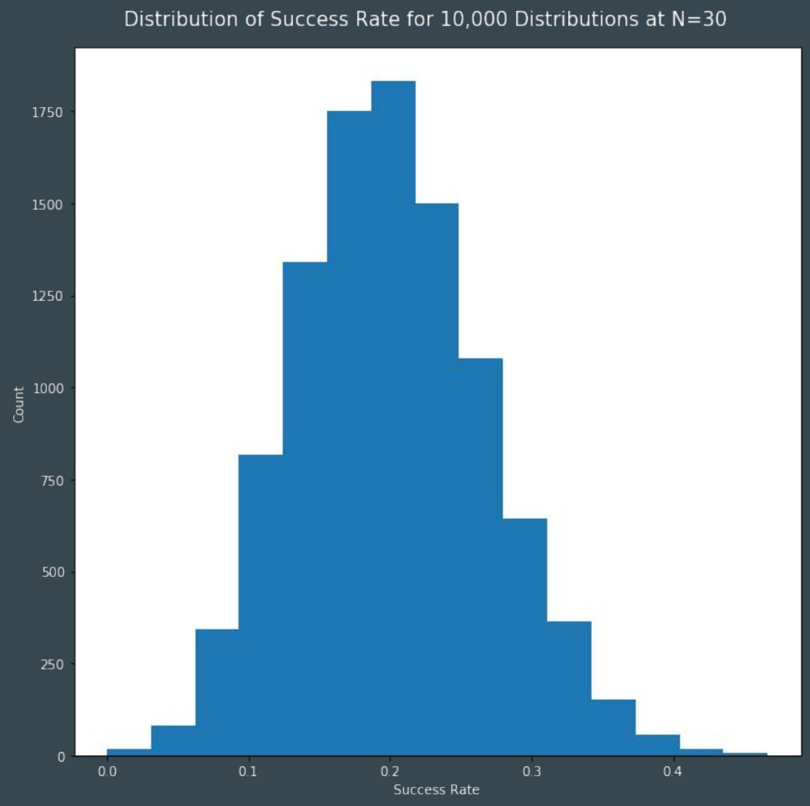
Estimation

- Say that I could call 30 customers in a given day. Every day I'd track the success rate of my calls.
- If I had 10,000 days of data, what would the resulting distribution of my success look like? What would the average be? What would the variability be?

	# of Calls	# of Successes	Success Rate
Day 1	30	4	0.133
Day 2	30	10	0.333
Day 3	30	6	0.2
Day 4	30	7	0.233

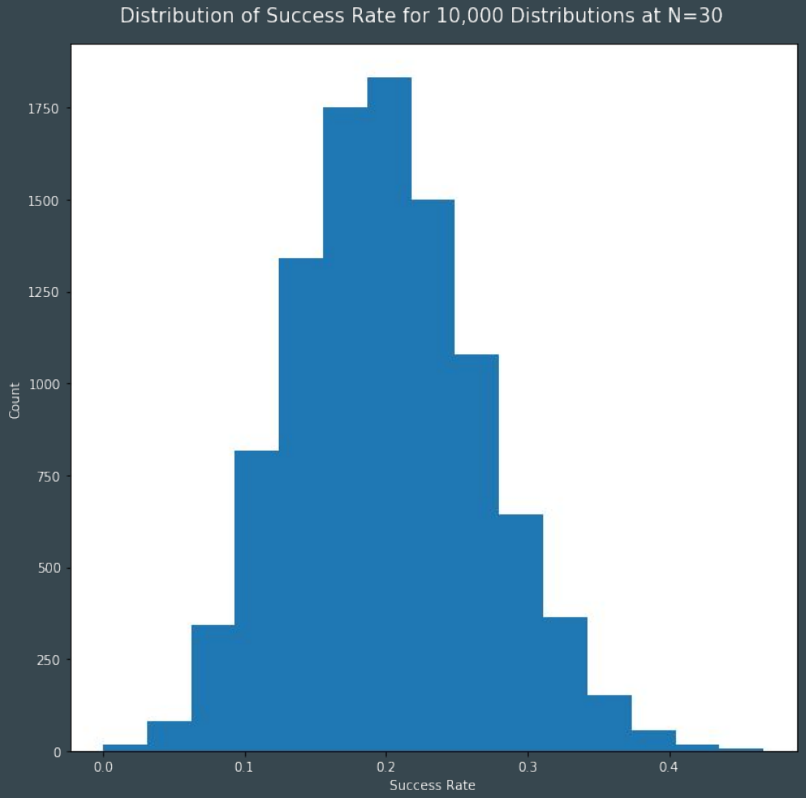
Estimation

- Note that I would have 10,000 data points, each containing the average success rate of 30 Bernoulli trials
- The resulting data points would look like a **normal distribution.**



Estimation

- The mean of my normal distribution here is the same as the underlying distribution, or 0.2.
- The standard deviation is my normal distribution is the standard deviation of the underlying distribution divided by the **square root** of the sample size.
- In this case, that is 0.16 divided by the square root of 30, approximately 0.073.



The Central Limit Theorem

- This phenomenon is called the **central limit theorem**.
- The **central limit theorem** states that if we take **repeated samples** from a distribution, the distributions of the **resulting sample means** will resemble a **normal distribution**.
- This normal distribution will have the **same mean** as the underlying distribution and a standard deviation equal to the **standard deviation from the underlying distribution divided by the square root of the sample size**.
- The most revelatory part of the central limit theorem is that it applies to **any distribution**, whether discrete or continuous. You could make up your own distribution and it would apply.
- The central limit theorem typically applies when there are at least **30 samples**, though this could be more or less depending on the underlying distribution.

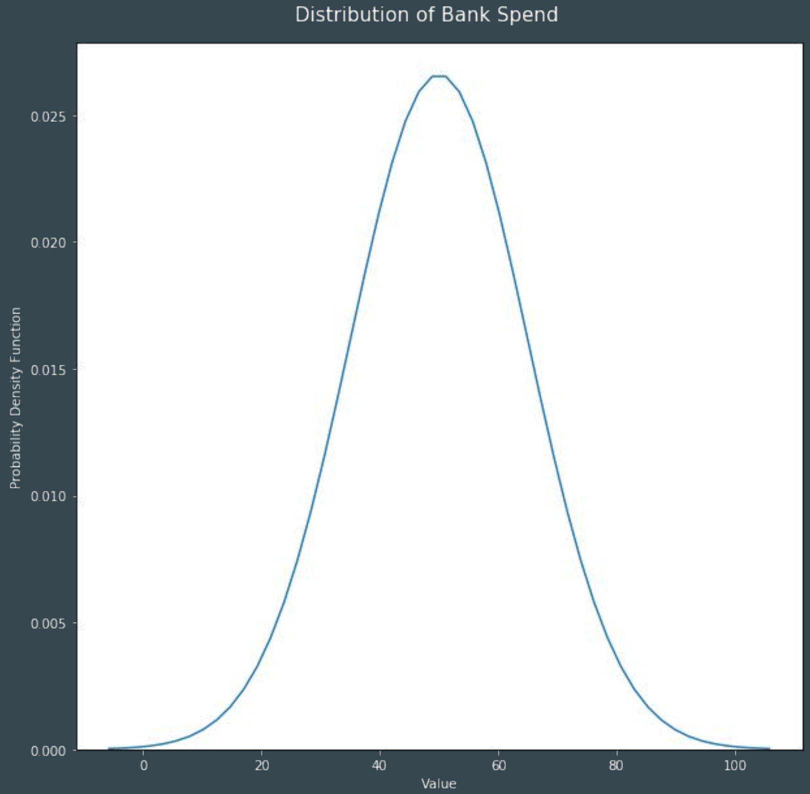
The Central Limit Theorem

- The most revelatory part of the central limit theorem is that it applies to **any distribution**, whether discrete or continuous. You could make up your own distribution and it would apply.
- The central limit theorem typically applies when there are at least **30 samples**, though this could be more or less depending on the underlying distribution.
- We will look at examples of this in the workbook portion.
- The **central limit theorem** is why the normal approximation works - since the binomial distribution is simply a repeated series of draws from a Bernoulli distribution, we could treat it the way we'd treat a normal distribution.

The Central Limit Theorem

- Let’s continue to look at my banking product, but at paying customers.
- Say that my customers tend to spend $50 per day in their banking account, with a standard deviation of $15 in a normal distribution.
- If I were to take 10,000 random samples of 100 customers, what would the distribution of my sample means look like? What would the mean be? What would the standard distribution be?

The Central Limit Theorem



Diff. in STD	Spend	Percentile
-3	$5	0.001
-2	$20	0.022
-1	$35	0.158
0	$50	0.5
1	$65	0.841
2	$80	0.977
3	$95	0.998

The Central Limit Theorem

- Let's continue to look at my banking product, but at paying customers.
- Say that my customers tend to spend $50 per day in their banking account, with a standard deviation of $15 in a normal distribution.
- If I were to take 10,000 random samples of 100 customers:
- What would the distribution of my sample means look like?
- What would the mean be?
- What would the standard distribution be?

The Central Limit Theorem

- Let's continue to look at my banking product, but at paying customers.
- Say that my customers tend to spend $50 per day in their banking account, with a standard deviation of $15 in a normal distribution.
- If I were to take 10,000 random samples of 100 customers:
- What would the distribution of my sample means look like?
  - *A normal distribution*
- What would the mean be?
- What would the standard distribution be?

The Central Limit Theorem

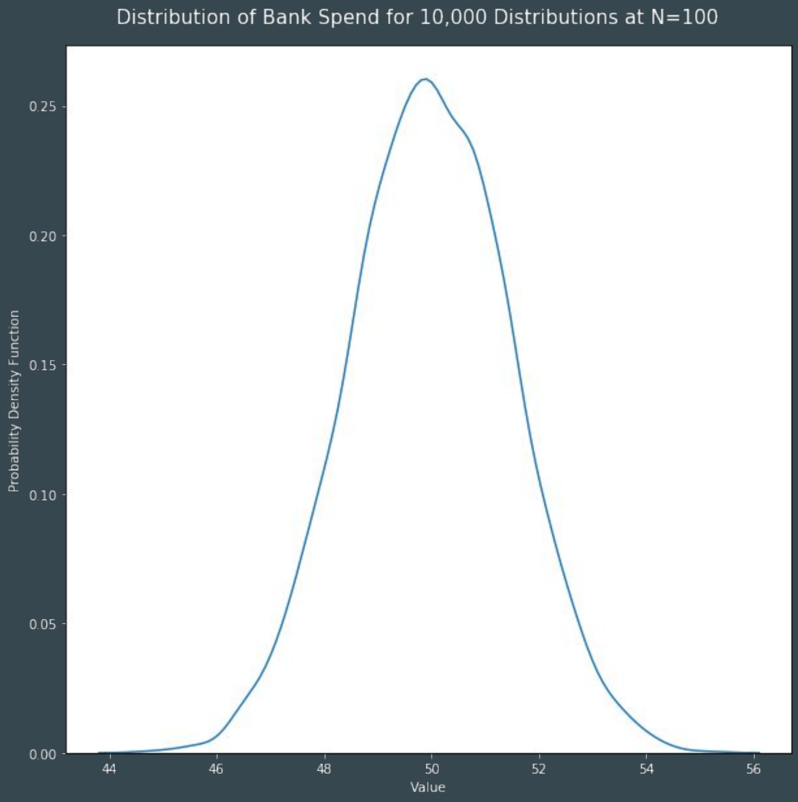
- Let’s continue to look at my banking product, but at paying customers.
- Say that my customers tend to spend $50 per day in their banking account, with a standard deviation of $15 in a normal distribution.
- If I were to take 10,000 random samples of 100 customers:
- What would the distribution of my sample means look like?
    - *A normal distribution*
- What would the mean be?
    - *$50*
- What would the standard distribution be?

The Central Limit Theorem

- Let's continue to look at my banking product, but at paying customers.
- Say that my customers tend to spend $50 per day in their banking account, with a standard deviation of $15 in a normal distribution.
- If I were to take 10,000 random samples of 100 customers:
- What would the distribution of my sample means look like?
  - *A normal distribution*
- What would the mean be?
  - *$50*
- What would the standard distribution be?
  - *$15 / 10 (the square root of 100), or $1.50*

Estimation

- My 10,000 trials would look like a **normal distribution.**
- The mean of my normal distribution is the same as the underlying distribution, or 0.2.
- The standard deviation is my normal distribution is the standard deviation of the underlying distribution divided by the **square root** of the sample size.
- In this case, that is 0.16 divided by the square root of 30, approximately 0.073.



The Central Limit Theorem

- Again, let's say we have an underlying normal distribution of distributions with a mean of \$50 and a standard deviation of \$15.
- In three scenarios:
  - A) We look at a single data point for a random customer.
  - B) We look at the sample mean from 20 different random customers.
  - C) We look at the sample mean from 100 different random customers.
- Which scenario do we expect a result closest to the mean of \$50?

The Central Limit Theorem

- Again, let's say we have an underlying normal distribution of distributions with a mean of $50 and a standard deviation of $15.
- In three scenarios:
  - A) We look at a single data point for a random customer.
  - B) We look at the sample mean from 20 different random customers.
  - **C) We look at the sample mean from 100 different random customers.**
- Which scenario do we expect a result closest to the mean of $50?
  - Scenario C - the more data points we have, the smaller our standard deviation will be, and thus the smaller chance our result will deviate from the mean.

The Central Limit Theorem

- The Central Limit Theorem is effective because of these two beliefs:
  - The mean of a random sample will be the same as the mean from the overall population
  - The standard error of a random sample will be smaller the greater you increase your sample size
- As we do **hypothesis testing** in the second half of the course, the central limit theorem will be a core tool in understanding whether a random sample is part of the distribution or not, regardless of what the underlying distribution is.

The Central Limit Theorem

- If I pull information from 100 paying customers and find that they are spending $60 on average, do we believe that they are similar to our existing customers, given what we know about the distribution of our paying customers? Is it statistically feasible?
- We know exactly what the distribution looks like for 100 paying customers and can do a **hypothesis test** on it to see how likely this outcome would be, given the rules of the **central limit theorem**
- Again, it will work on **any distribution**, generally with only 30 samples needed, any many times less.