Week Seven: Hypothesis Testing

CS 217

Introduction

- Say we want to see if a coin is unfair or not. If we toss it 10 times, how many heads would make us think it was unfair?
  - 9 Heads?
  - 0 Heads?
  - 3 Heads?

Introduction

- Say we want to see if a coin is unfair or not. If we toss it 10 times, how many heads would make us think it was unfair?
  - 9 Heads? Probably. Probably?
  - 0 Heads? Definitely? Definitely?
  - 3 Heads? Maybe? Could just be chance?

Introduction

- Say we want to see if a coin is unfair or not. If we toss it 10 times, how many heads would make us think it was unfair?
  - 9 Heads? Probably. Probably?
  - 0 Heads? Definitely? Definitely?
  - 3 Heads? Maybe? Could just be chance?
- **Hypothesis Testing** is a way of thinking about this probabilistically.

Definitions

- **H_0 - Null Hypothesis** - The default assumption for the model generating the data
  - Everything is fine!
- **H_A - Alternative Hypothesis** - The alternate explanation for our data given that we reject the null hypothesis
  - Something is weird, and not as originally assumed.

Definitions

- **H_0 - Null Hypothesis** - The default assumption for the model generating the data
- **H_A - Alternative Hypothesis** - The alternate explanation for our data given that we reject the null hypothesis

- **H_0 - Null Hypothesis** - Our coin is fair (has a 50% chance of heads and tails)
- **H_A - Alternative Hypothesis** - Our coin is not fair (has something different than a 50% chance of heads and tails)
- Note that at the moment we are not concerned about what other distribution the coin may have, just whether it fits our null hypothesis or not.

Definitions

- **X - Test Statistic** - The result we observe in our experiment.
- **Null distribution** - the probability distribution of X assuming H_0

  What distribution does a coin flipped ten times follow?

Definitions

- We assume our coin follows a **binomial distribution** with **ten trials** and a probability of success per trial of **0.5**.

Head Count	PMF	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Definitions

- **Rejection region** - if X is in the rejection region, we reject H_0 in favor of H_A
- **Non-Rejection region** - the **complement** to the rejection region, if X is in this region we do not reject H_0

Definitions

- **Rejection region** - if X is in the rejection region, we reject H_0 in favor of H_A
- **Non-Rejection region** - the **complement** to the rejection region, if X is in this region we do not reject H_0

- **Rejection region** - The number of heads in 10 flips that seems suspicious to us.
- **Non-Rejection region** - The number of heads in 10 flips that do not seem suspicious to us.

How do we determine what is ‘suspicious’ to us?

What's a Fair Coin, Anyway?

- We know we can use the PMF function to find the odds of getting a specific number of heads in ten flips.
- From there, it's up to us to determine what we deem as suspicious.

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

What's a Fair Coin, Anyway?

- We know we can use the PMF function to find the odds of getting a specific number of heads in ten flips.
- From there, it's up to us to determine what we deem as suspicious
- For example, we can say that 2 or less heads in 10 coin flips or 8 or greater heads in 10 coin flips is 'suspicious'
- Given the PMFs of the distribution, there is an 11% chance of this occurring if the coin is fair

Head Count	Odds
0	0.001
1	0.01
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.01
10	0.001

Definitions

- **H_0 - Null Hypothesis** - The coin that we're flipping is fair.
- **H_A - Alternative Hypothesis** - The coin we're flipping is not fair.
- **X - Test Statistic** - We see _ heads in 10 flips.
- **Null distribution** - The null distribution is a binomial distribution with 10 trials and a probability of success of 50% per trial.
- **Rejection region** - Getting 0, 1, 2, 8, 9, or 10 heads in 10 coin flips
- **Non-Rejection region** - Getting 3, 4, 5, 6, or 7 heads in 10 coin flips

Definitions

- **H_0 - Null Hypothesis** - The coin that we're flipping is fair.
- **H_A - Alternative Hypothesis** - The coin we're flipping is not fair.
- **X - Test Statistic** - We see _ heads in 10 flips.
- **Null distribution** - The null distribution is a binomial distribution with 10 trials and a probability of success of 50% per trial.
- **Rejection region** - Getting 0, 1, 2, 8, 9, or 10 heads in 10 coin flips
- **Non-Rejection region** - Getting 3, 4, 5, 6, or 7 heads in 10 coin flips

- The probability of us rejecting **H_0 given that H_0 is true is 11%**
- There is an 11% chance that, given the null distribution, we will end up in the rejection region.

Definitions

- Given that there are two possibilities for what the coin is (the coin is fair or the coin is not fair), and two possibilities for how we respond to the experiment (we reject the null hypothesis or we *fail to reject* the null hypothesis), there are four outcomes that can occur:
- The coin is fair and we fail to reject the null hypothesis
  - True Negative
- The coin is fair and we reject the null hypothesis
  - False Positive - Type I error
- The coin is not fair and we fail to reject the null hypothesis
  - False Positive - Type II error
- The coin is not fair and we reject the null hypothesis
  - True Positive

Definitions

	Coin is fair (H_0)	**Coin is not fair (H_A)**
Reject H_0	Type I Error - False Positive	True Positive
Don't Reject H_0	True Negative	Type II Error - False Negative

Think of '**positive**' as **rejecting H_0** and **'negative'** as **failing to reject H_0**

Definitions

- The probability of us **rejecting H_0 given that H_0 is true** is 11%.
- Rejecting H_0 given that H_0 is true is a Type I error, or a false positive.
- In hypothesis testing, this is also known as the **significance level**, **alpha,** or **p-value.**
- Note that in this example we picked out a rejection region manually and backed into a significance level, but typically we pick a significance level and see which results will fall into a rejection region.

Definitions

- The probability of us **rejecting H_0 given that H_0 is true** is 11%.
- Rejecting H_0 given that H_0 is true is a Type I error, or a false positive.
- In hypothesis testing, this is also known as the **significance level**, **alpha,** or **p-value.**
- Note that in this example we picked out a rejection region manually and backed into a significance level, but typically we pick a significance level and see which results will fall into a rejection region.

- Failing to reject H_0 given that it is false is a Type II error, or a false negative.
- In hypothesis testing, the **inverse of this** is known as the **power level** or **beta** value.
- We want both the **alpha** value to be as low as possible and the **beta** value to be as high as possible when conducting our test.

Example

- Let's say we are testing a new cancer drug.
- H_0 is the null hypothesis, that this drug is no more effective than a placebo.
- H_A is the alternate hypothesis, that this drug is more effective than a placebo.

Example

- Let's say we are testing a new cancer drug.
- H_0 is the null hypothesis, that this drug is no more effective than a placebo.
- H_A is the alternate hypothesis, that this drug is more effective than a placebo.
- A false positive is when we **reject the null hypothesis** given the results of our experiment even though the null hypothesis is true.
- Here, this is that we say the drug is more effective than a placebo, even though it actually isn't.
- The **significance level** is the probability that this will occur given that the null hypothesis is true.

Example

- Let's say we are testing a new cancer drug.
- H_0 is the null hypothesis, that this drug is no more effective than a placebo.
- H_A is the alternate hypothesis, that this drug is more effective than a placebo.
- A false negative is when we **fail to reject the null hypothesis** given the results of our experiment even though the **null hypothesis is not true.**
- Here, this is that we say the drug is not more effective than a placebo, even though it actually is.
- The **power level** is the inverse probability of this occurring (given that the alternative hypothesis is true).
- Again, we want both the **alpha** value to be as low as possible and the **beta** value to be as high as possible when conducting our test.

Example

- Let's say we are trying someone in court for murder.
- H_0 is the null hypothesis, that this person did not commit the crime.
- H_A is the alternate hypothesis, that this person did commit the crime.

Example

- Let's say we are trying someone in court for murder.
- H_0 is the null hypothesis, that this person did not commit the crime.
- H_A is the alternate hypothesis, that this person did commit the crime.
- A false positive is when we **reject the null hypothesis** given the results of our experiment even though the null hypothesis is true.
- Here, this is that we say the person did commit the crime, even though they actually didn't.
- The **significance level** is the probability that this will occur.

Example

- Let's say we are trying someone in court for murder.
- H_0 is the null hypothesis, that this person did not commit the crime.
- H_A is the alternate hypothesis, that this person did commit the crime.
- A false negative is when we **fail to reject the null hypothesis** given the results of our experiment even though the **null hypothesis is not true.**
- Here, this is that we say the person did not commit the crime, even though they actually did.
- The **power level** is the inverse probability of this occurring.
- Again, we want both the **alpha** value to be as low as possible and the **beta** value to be as high as possible when conducting our test.

Hypothesis Test Design

- Let's say again that we're examining whether a coin is unfair.
- 1. Pick the null hypothesis H_0
  - Here our null hypothesis is that 'the coin is not biased', or that specifically the probability of landing heads on a given coin flip is 0.5
- 2. Decide if H_A is one-sided or two-sided
  - Do we only care if the coin is biased towards heads (we get 8, 9, 10 coin flips)? Or biased towards heads or tails (we get 0, 1, 2 coin flips)?
- 3. Pick a significance level and determine the rejection region
  - Typical significance levels include 0.1, 0.05, and 0.01.
  - The significance level is equal to the probability of a false positive given the null hypothesis

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - **P(H) = 0.5**
- 2. Decide if H_A is one-sided or two-sided
  - Two-sided
- 3. Pick a significance level and determine the rejection region
  - 0.01

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - P(H) = 0.5
- 2. Decide if H_A is one-sided or two-sided
  - Two-sided
- 3. Pick a significance level and determine the rejection region
  - 0.01

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - P(H) = 0.5
- 2. Decide if H_A is one-sided or two-sided
  - Two-sided
- 3. Pick a significance level and determine the rejection region
  - 0.05

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - P(H) = 0.5
- 2. Decide if H_A is one-sided or two-sided
  - Two-sided
- 3. Pick a significance level and determine the rejection region
  - 0.05

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - P(H) = 0.5
- 2. Decide if H_A is one-sided or two-sided
  - One-sided, we only care about coins that are biased towards heads
- 3. Pick a significance level and determine the rejection region
  - 0.1

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - P(H) = 0.5
- 2. Decide if H_A is one-sided or two-sided
  - One-sided, we only care about coins that are biased towards heads
- 3. Pick a significance level and determine the rejection region
  - 0.1

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - P(H) = 0.5
- 2. Decide if H_A is one-sided or two-sided
  - One-sided, we only care about coins that are biased towards tails
- 3. Pick a significance level and determine the rejection region
  - 0.1

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

Hypothesis Test Design

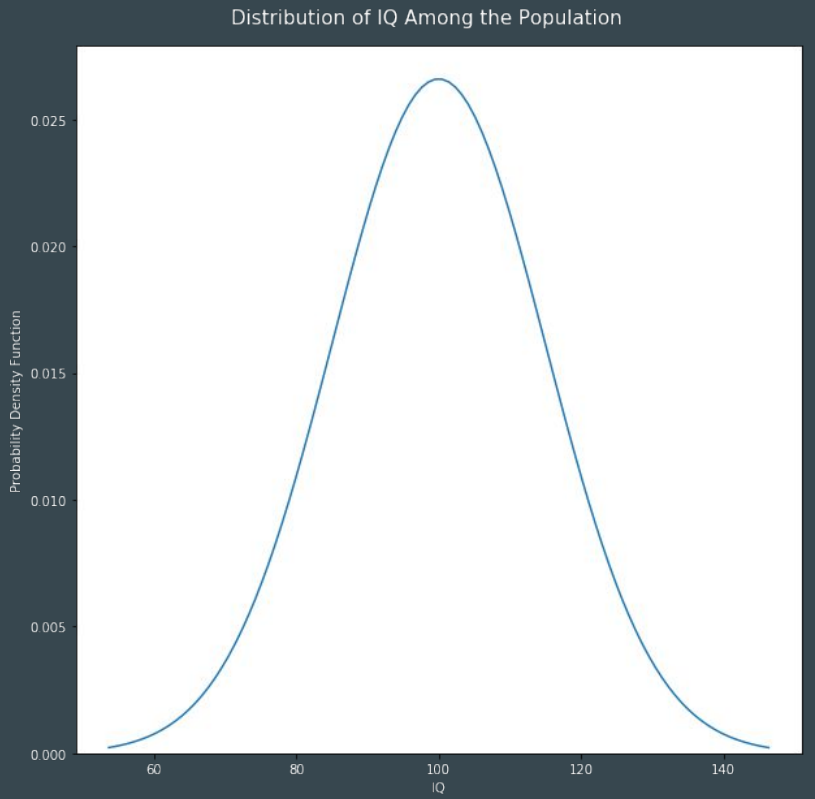
- 1. Pick the null hypothesis H_0
  - P(H) = 0.5
- 2. Decide if H_A is one-sided or two-sided
  - One-sided, we only care about coins that are biased towards tails
- 3. Pick a significance level and determine the rejection region
  - 0.1

Which results are in our rejection region in this case?

Head Count	Odds	5	0.246
0	0.001	6	0.205
1	0.01	7	0.117
2	0.044	8	0.044
3	0.117	9	0.01
4	0.205	10	0.001

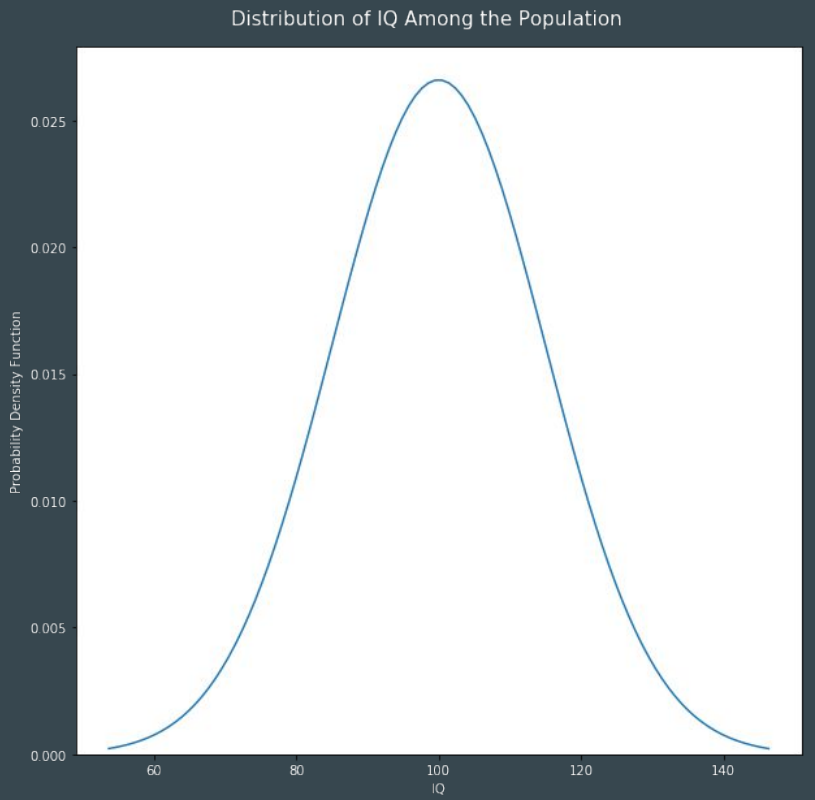
Hypothesis Test Design

- IQ is normally distributed in the population with a mean of 100 and a standard deviation of 15.
- Say we think that CCNY students have, on average, an above average intelligence.
- What might our null hypothesis be here?
- What might our alternative hypothesis be?



Hypothesis Test Design

- IQ is normally distributed in the population with a mean of 100 and a standard deviation of 15.
- Say we think that CCNY students have, on average, an above average intelligence.
- What might our null hypothesis be here?
  - CCNY students have, on average, an IQ of 100
- What might our alternative hypothesis be?
  - CCNY students have, on average, an IQ greater than 100



Hypothesis Test Design

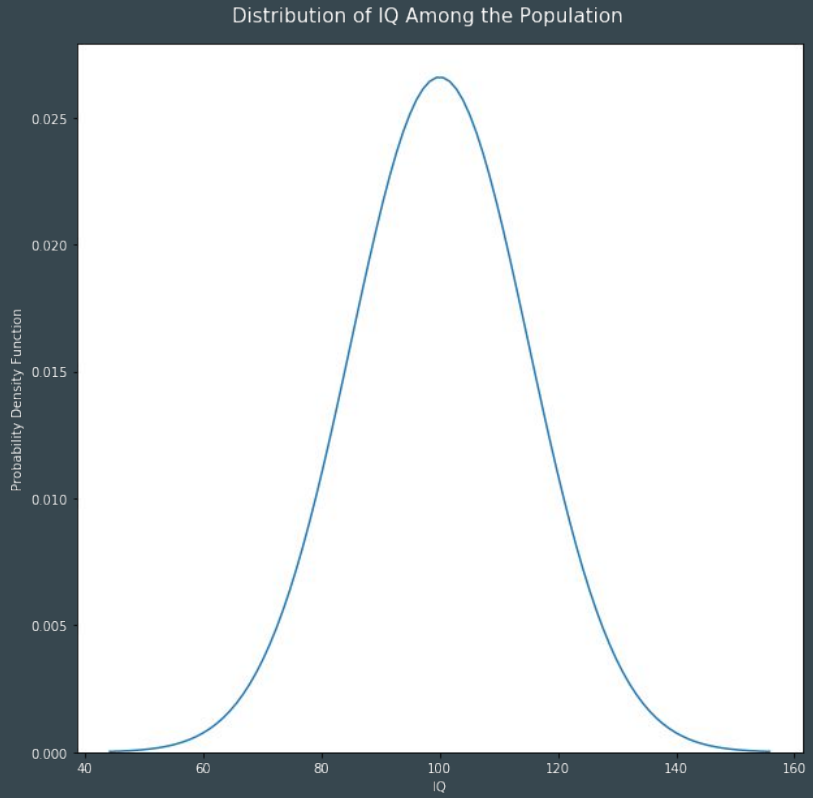
- 1. Pick the null hypothesis H_0
- 2. Decide if H_A is one-sided or two-sided
- 3. Pick a significance level and determine the rejection region

Hypothesis Test Design

- 1. Pick the null hypothesis H_0
  - The mean of IQ of CCNY students is 100
- 2. Decide if H_A is one-sided or two-sided
  - We only care if the mean IQ of CCNY students is greater than 100
  - Thus, our test is one-sided
- 3. Pick a significance level and determine the rejection region
  - Let's use a significance level of 0.05.
  - You can also determine a specific rejection region like we did earlier, but using a significance level is much more common.
- Let's say we ask nine students their IQ and take the mean of that. How do we determine the rejection region?
  - We will reject the null hypothesis if the nine people we poll have a mean IQ greater than the 95th (1 - 0.05) percentile of our distribution.

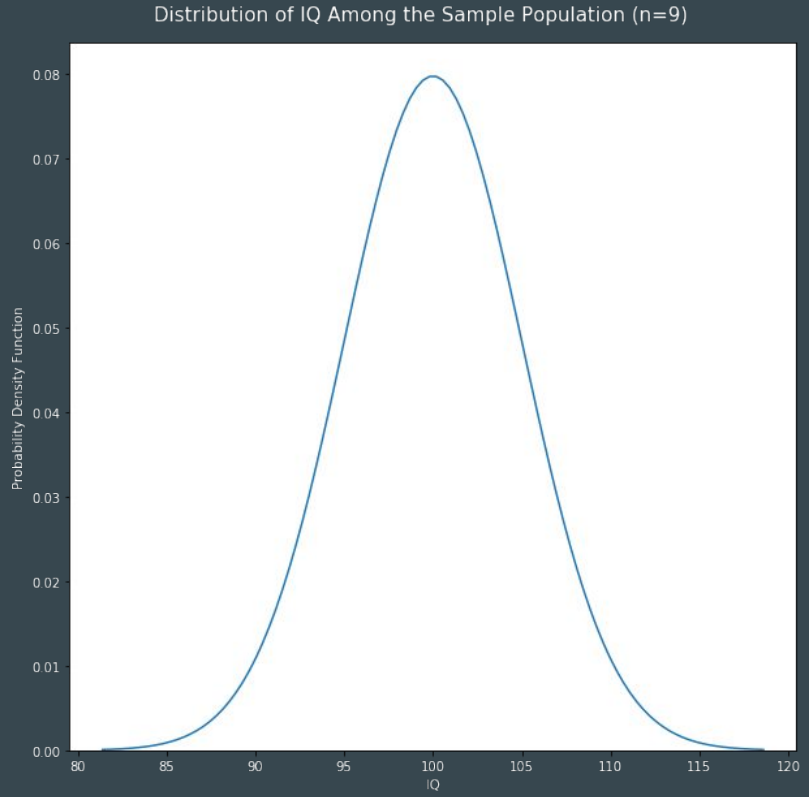
Hypothesis Test Design

- Like we discussed, a few weeks ago, the distribution of the mean IQ for nine people is different from the distribution of IQ for the population.
- The population IQ has a mean of 100 and a standard deviation of 15.
- What does the distribution of the mean IQ for a sample of 9 people have a mean and standard deviation of?



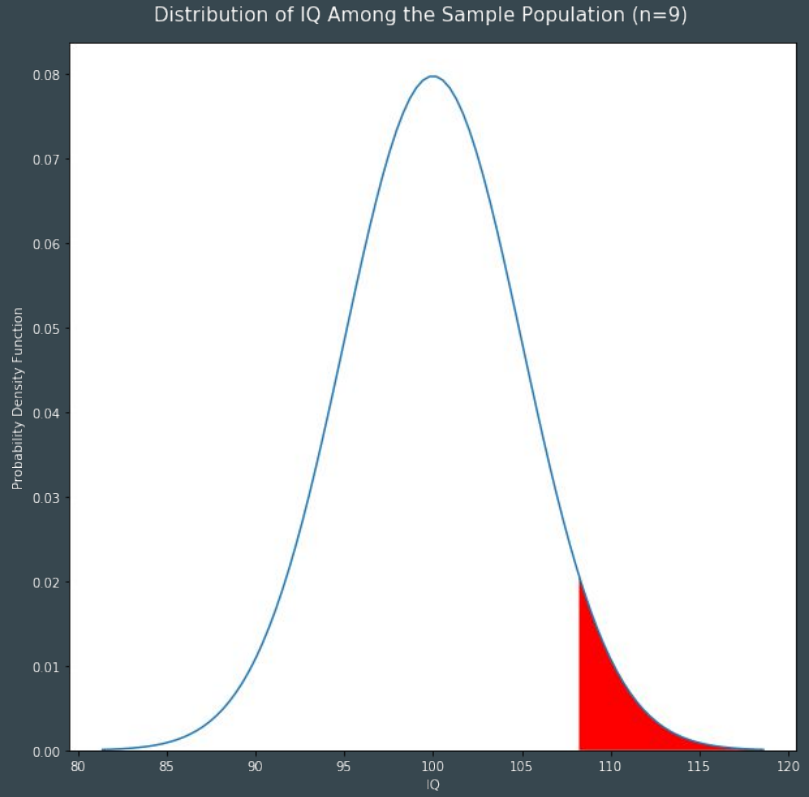
Hypothesis Test Design

- Like we discussed, a few weeks ago, the distribution of the mean IQ for nine people is different from the distribution of IQ for the population.
- The population IQ has a mean of 100 and a standard deviation of 15.
- What does the distribution of the mean IQ for a sample of 9 people have a mean and standard deviation of?
  - Mean = 100
  - STD = 15 / sqrt(9) = 5



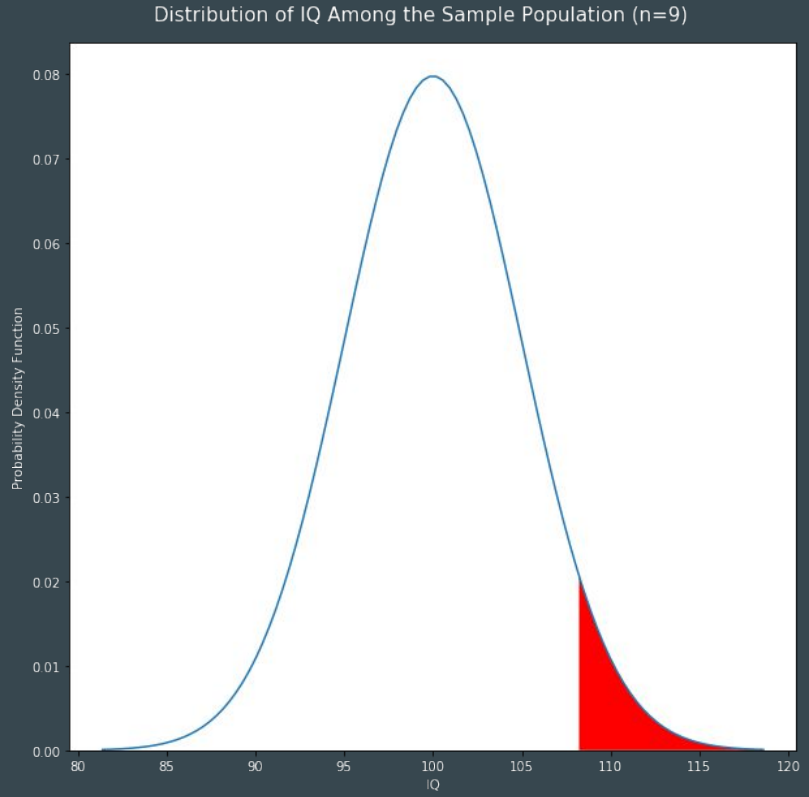
Hypothesis Test Design

- Similar to what we did for the discrete distribution, we will **reject** the null hypothesis for any results that come in above the 95th (1 - 0.05) percentile.
- Specifically that means anything **greater than or equal to** a mean IQ of 108.22 for the nine people we speak with



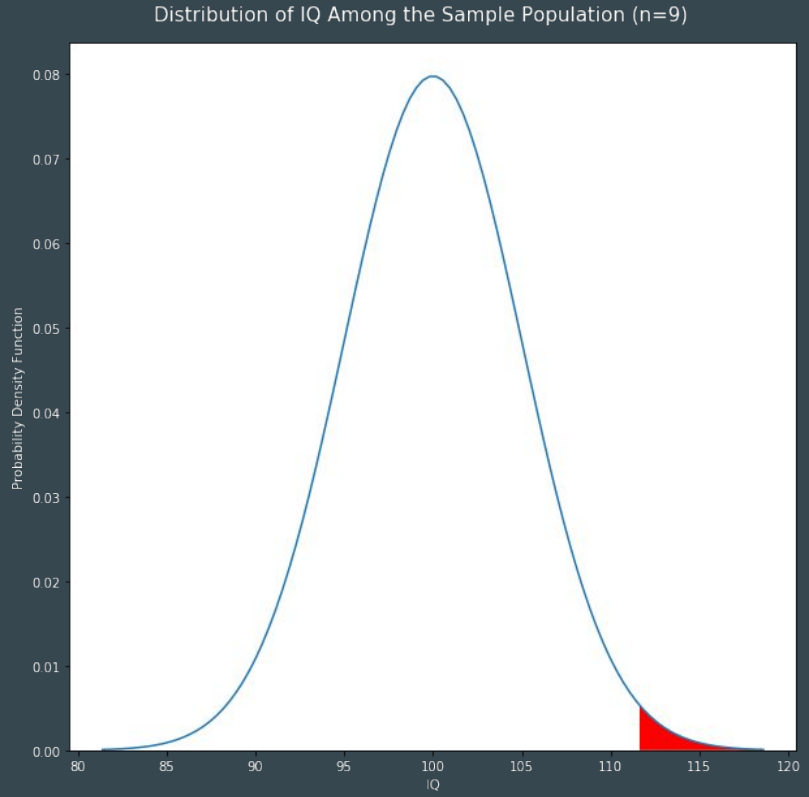
Hypothesis Test Design

- If our nine students have a mean IQ of 110, we can **reject the null hypothesis** that the mean IQ of CCNY students is equal to 100
- If our nine students have a mean IQ of 105, we fail to reject the null hypothesis that the mean IQ of CCNY students is equal to 100
- Remember that our significance level is 0.05. What if we decrease it to 0.01?



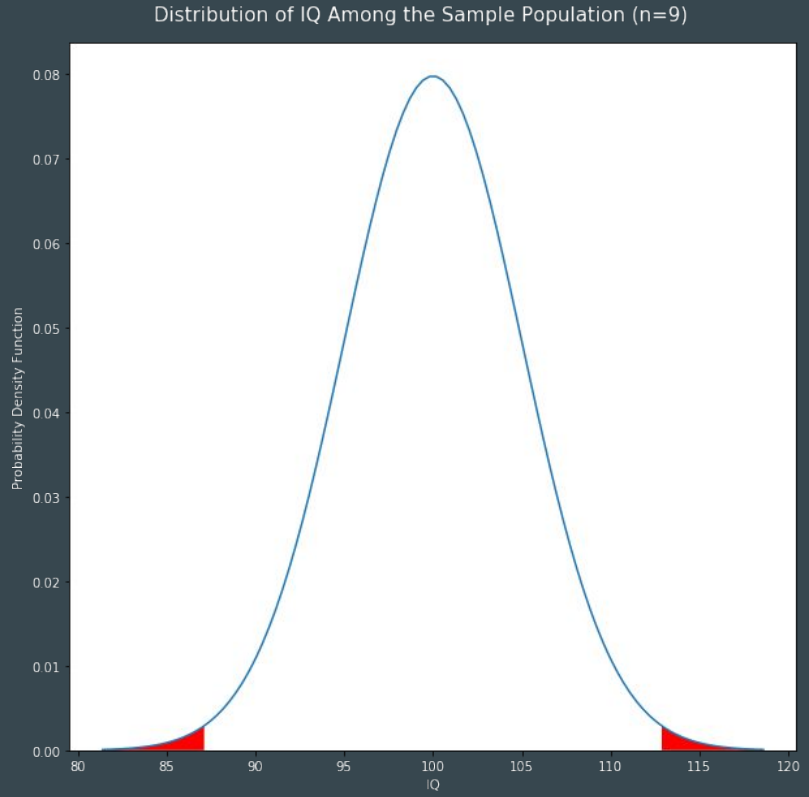
Hypothesis Test Design

- Now we will reject any results that are above the 99th percentile.
- Specifically that means a mean IQ of 111.63 for the nine people we speak with
- If our nine students have a mean IQ of 110, we now **reject the null hypothesis** that the mean IQ of CCNY students is equal to 100



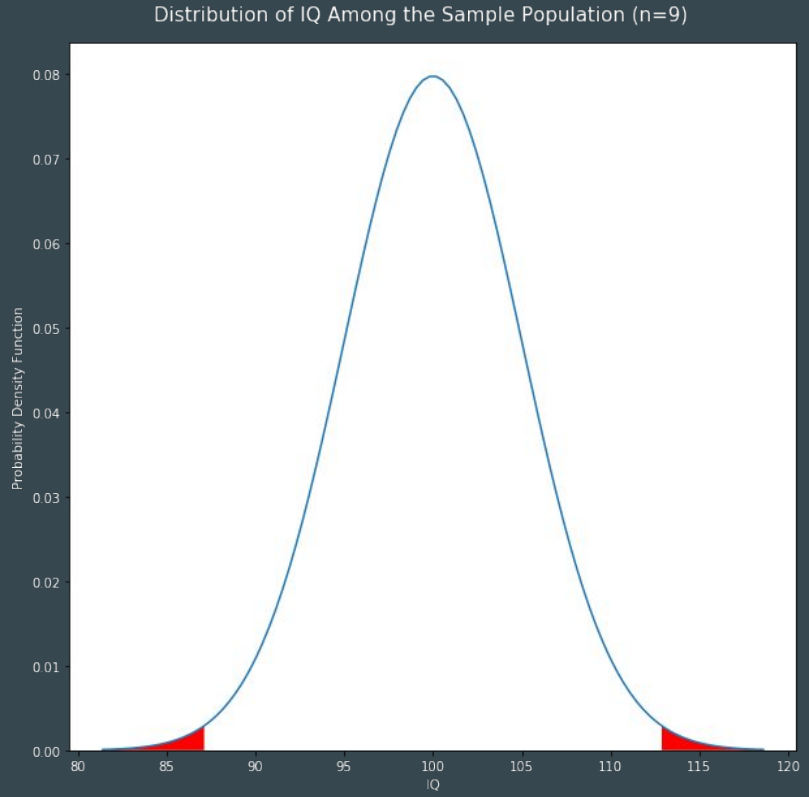
Hypothesis Test Design

- Now let's do a two-sided test rather than a one-sided test.
- We are now testing whether the mean IQ of a CCNY is *different* than the mean IQ of the population, rather than greater.
- Now, with a significance level still at 0.01, the rejection region consists of the area before the 0.5th (half of the 1st) percentile and greater than the 99.5th percentile.



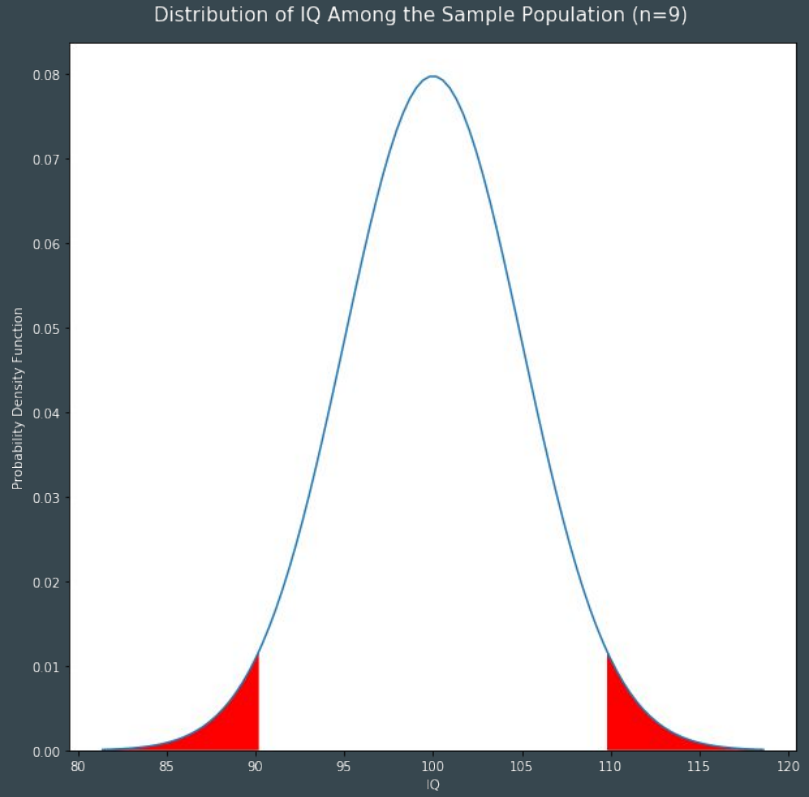
Hypothesis Test Design

- This is equivalent to an IQ less than 87.12, or greater than 112.87.
- If our nine students have a mean IQ of 110, we will *fail to reject* the null hypothesis that the mean IQ of CCNY students is the same as the mean IQ of the general population.



Hypothesis Test Design

- Now let's change the significance level from 0.01 to 0.05.
- The rejection region will now encompass anywhere between the 2.5th percentile and the 97.5th percentile.
- This is equivalent to an IQ less than 90.2 or an IQ greater than 109.79
- Now we will *reject* the null hypothesis



Levers to Pull

- Now that we have seen a few permutations of different results, what are the levers we know we can pull?
- Changing the number of people we poll
  - This will change our null distribution, as the standard deviation will be affected by the number of people in our sample.
- Changing between a one-sided and two-sided test
  - This will change our rejection region.
- Changing our significance level
  - The most common significance level is 0.05, but 0.01 and 0.1 (rarely) are also used. What is our tolerance for a false positive?
- Choosing these parameters are often more **art** than **science** (though evaluating a test within a set of chosen parameters is a science!!)

Estimation

- Thus far, we have worked with data where we have a known population mean and population variance
- In our IQ example, we know that the population mean is 100 and the population variance is the square of the population standard deviation, or $15^2 = 225$.
- In our coin flip example, we know that the population mean is 5 and the population variance is 1.25
- But what if we don't know the population variance? We can *estimate* it using our sample data.

Estimation

- We previously knew that the nine students at CCNY had an average IQ of 110.
- To the right are the nine student samples that we took.
- The **sample mean** of our nine samples is 110
- What about the sample variance?

1. 112	6. 91
2. 94	7. 142
3. 116	8. 119
4. 140	9. 85
5. 91	

Estimation

- Traditionally the variance entails taking the sum of the squared deviations of the dataset, divided by the length of the dataset.
- Here, the sum of squared deviations is 3648. If we divide that by the length of the dataset (9), that gives us a variance of 405.33
- The square root of 405.33 is about 20.12, which would be our standard deviation.

Value	Mean	Deviation	Sq. Deviation
112	110	2	4
94	110	-6	36
116	110	6	36
140	110	30	900
91	110	-9	81
91	110	-9	91
142	110	32	1024
119	110	9	81
85	110	-15	225

Estimation

- However, the **sample** variance entails taking the sum of the squared deviations of the dataset, divided by the length of the dataset **minus one**.
- Here, the sum of squared deviations is 3648. If we divide that by the length of the dataset - 1 (8), that gives us a **sample variance** of 456.
- The square root of 456 is about 21.35, which would be our **sample standard deviation.**

Value	Mean	Deviation	Sq. Deviation
112	110	2	4
94	110	-6	36
116	110	6	36
140	110	30	900
91	110	-9	81
91	110	-9	91
142	110	32	1024
119	110	9	81
85	110	-15	225

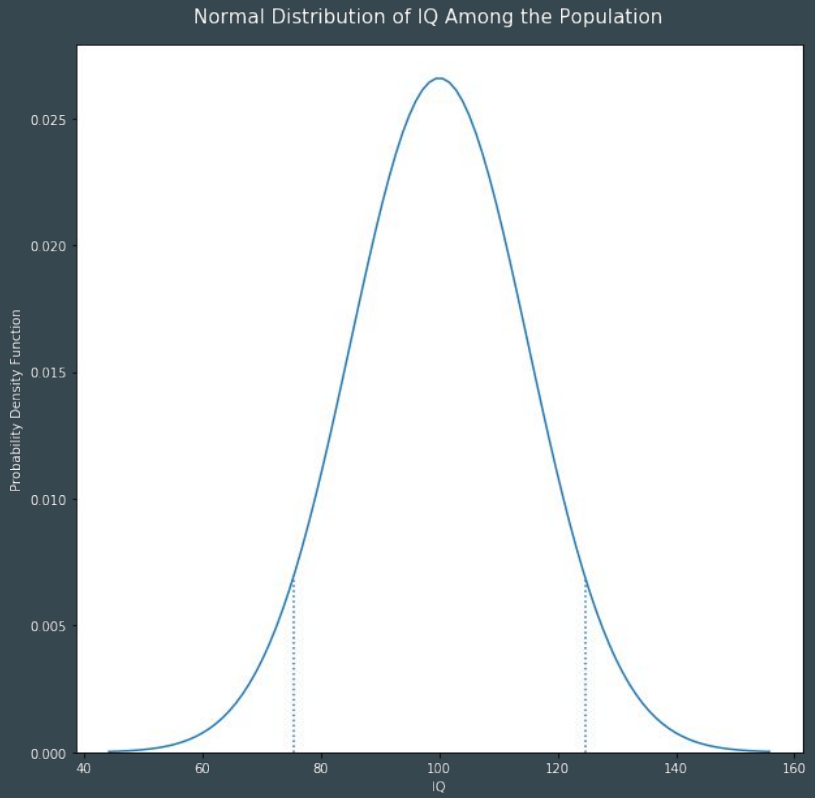
T Distribution

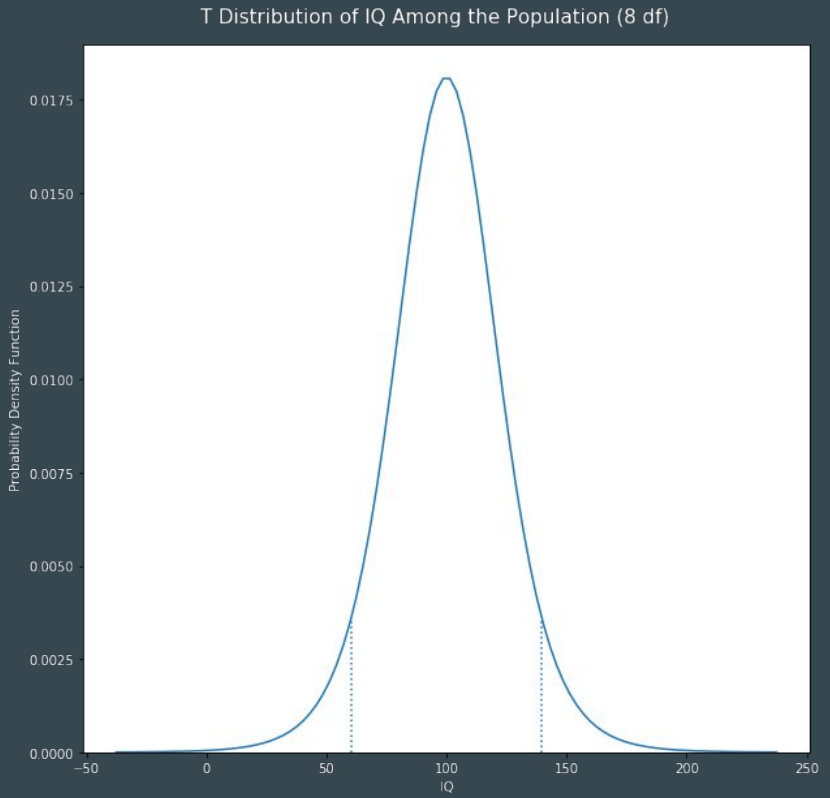
- Now, given our samples, we want to see if we can reject the original null hypothesis that the mean IQ for CCNY students is equal to 100 at a significance level of 0.05
- Originally, we had a **normal distribution** for the null hypothesis, with a mean of 100 and a standard deviation of 15.
- Now, we will use a **T-Distribution** for the null hypothesis, with a mean of 100 and standard deviation of 21.35.
- The **T-Distribution** is a bell-shaped curve, similar to the normal distribution, that has more probability in its tails than the standard normal distribution (due to the uncertainty about the true population variance)
- The T-Distribution changes based on the number of samples its based on. This metric is called *degrees of freedom*.

T Distribution

- The *degrees of freedom* are equal to the number of samples minus 1.
- In our case that will be 8, since there were 9 samples.

T Distribution

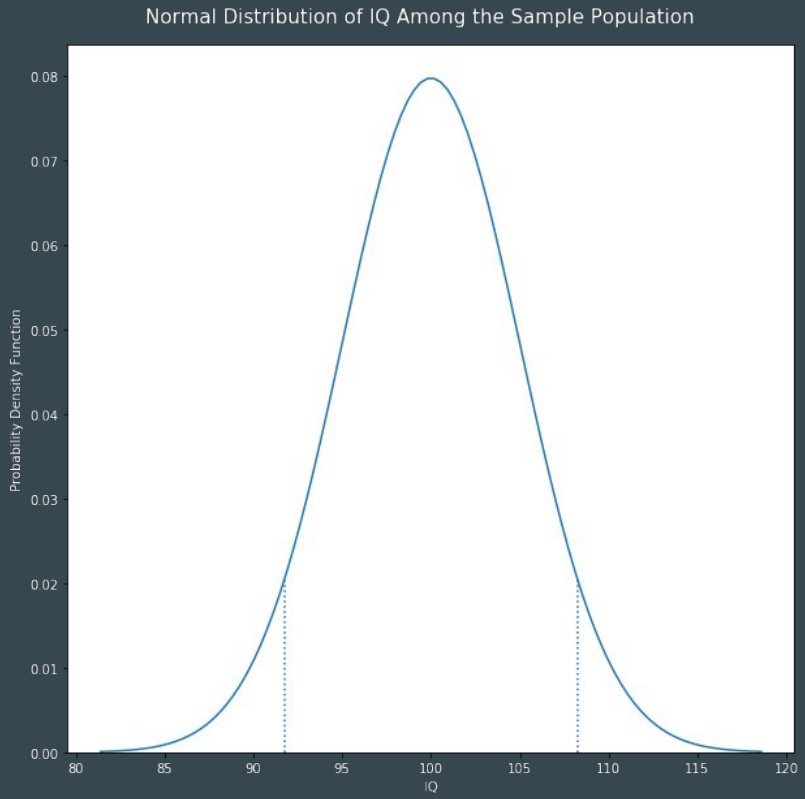


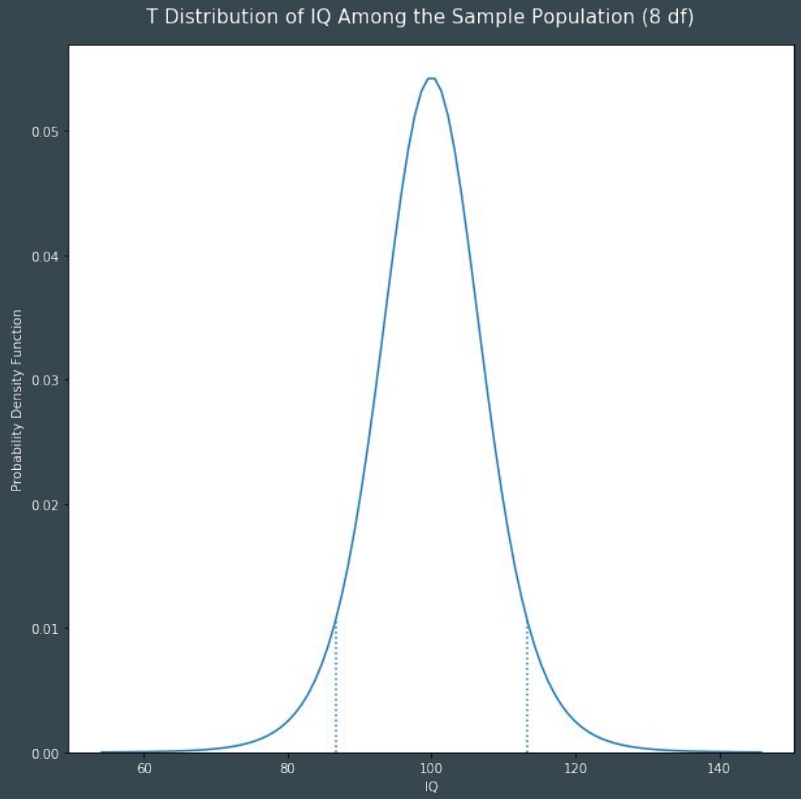


T Distribution

- The *degrees of freedom* are equal to the number of samples minus 1.
- In our case that will be 8, since there were 9 samples.
- Note that the rejection region is **much more extreme** for our T-distribution than our normal distribution. This is both because the sample variance is higher than the original population variance, but also because the T-distribution will inherently have a wider distribution than the normal distribution, even if the variance were the same.
- Let's now look at our *sample distributions*, given that we are looking at a sample of 9 people in both situations.

T Distribution





T Distribution

- Now we can follow our standard procedure and create **rejection regions**, given a two-sided test at a significance level of 0.05
- We will reject any values with an IQ less than 83.58 or greater than 116.41
- This is compared to a normal distribution, where we'll reject any values with an IQ less than 86.04 or greater than 113.95.
- In either case, we *fail to reject* the null hypothesis that the mean IQ of the CCNY student (at 110) is higher (*or different)* than the mean IQ of the general population.