Week Nine: Hypothesis Testing II

•••

CSC 217

Recap

- Last week, we went over a few different scenarios where a hypothesis test would be useful.
- In both scenarios, we want to test the mean of a series of samples versus a hypothesized population mean.
- We use a **z-test when** we know the variance of the underlying distribution.
- We use the t-test when we do not the variance of the underlying distribution.

Z-Test

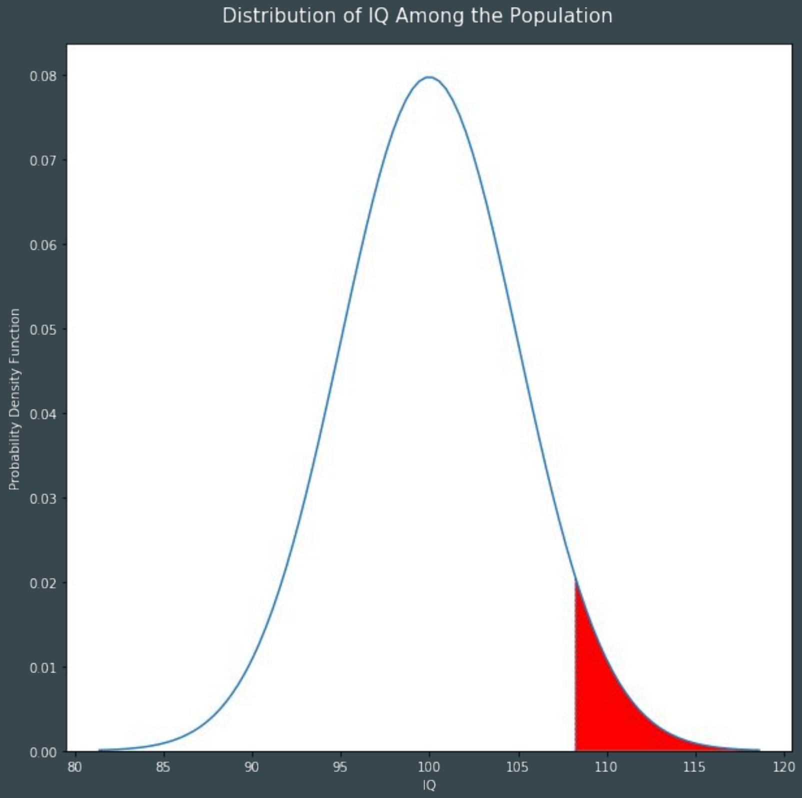
- With a z-test, we are testing if the population mean equals the hypothesized mean.
- The data we are referencing are **independent normal samples** from the distribution. We know the inherent variance of the distribution.
- The null hypothesis is that the population mean is equal to the hypothesis mean.
- The alternative hypothesis can be different depending on the type of test we're doing.
- For a two-sided test, the alternate hypothesis is that the population mean is not equal to the hypothesized mean
- For a one-sided test, the alternate hypothesis could be that hypothesized mean is greater than the population mean
- Or it could be that the hypothesized mean is less than the population mean

Z-Test

- We want to test if the mean IQ of the CCNY student is **greater than** the mean population IQ, given that IQ is normally distributed in the population with a mean of 100 and a standard deviation of 15.
- We take nine samples of CCNY students and find that they (you) have a mean IQ of 112.
- Given these parameters here - we want to do a one-sided test, we know the inherent variance of the population and we are sampling nine people, we can do a specific z-test with these parameters.

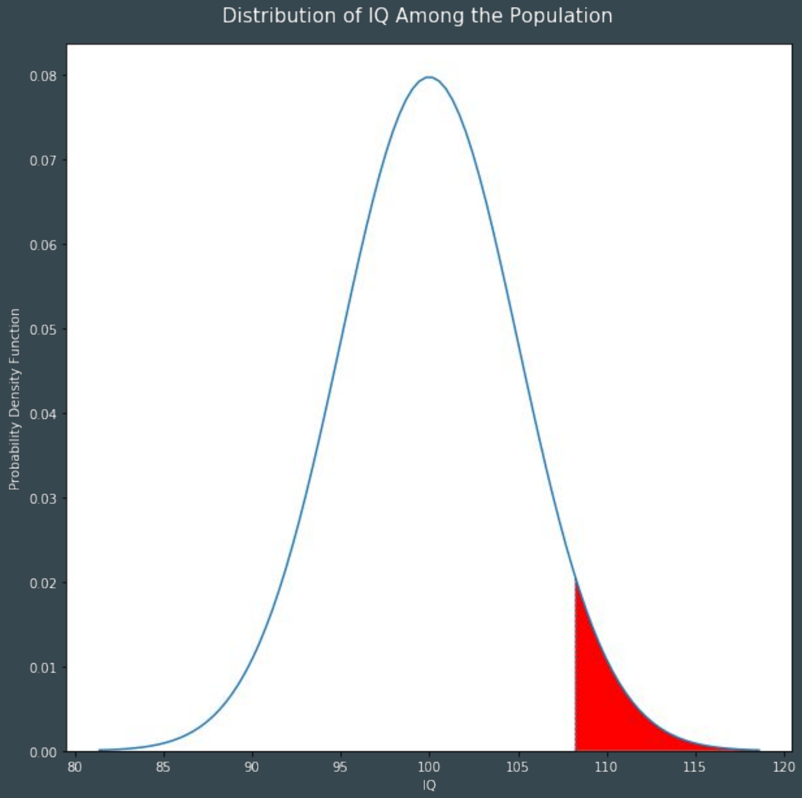
Z-Test

- We can set up a normal distribution with a mean of 100 and a standard deviation of 5, or the standard deviation (15) of the population divided by the square root of our sample size (9).
- The rejection region will be anything greater than the 95th percentile of this distribution, or 108.22.



Z-Test

- Here, we will reject the null hypothesis that CCNY students have the same mean IQ as the general population in favor of the alternate hypothesis that CCNY students have a greater mean IQ than the general population.

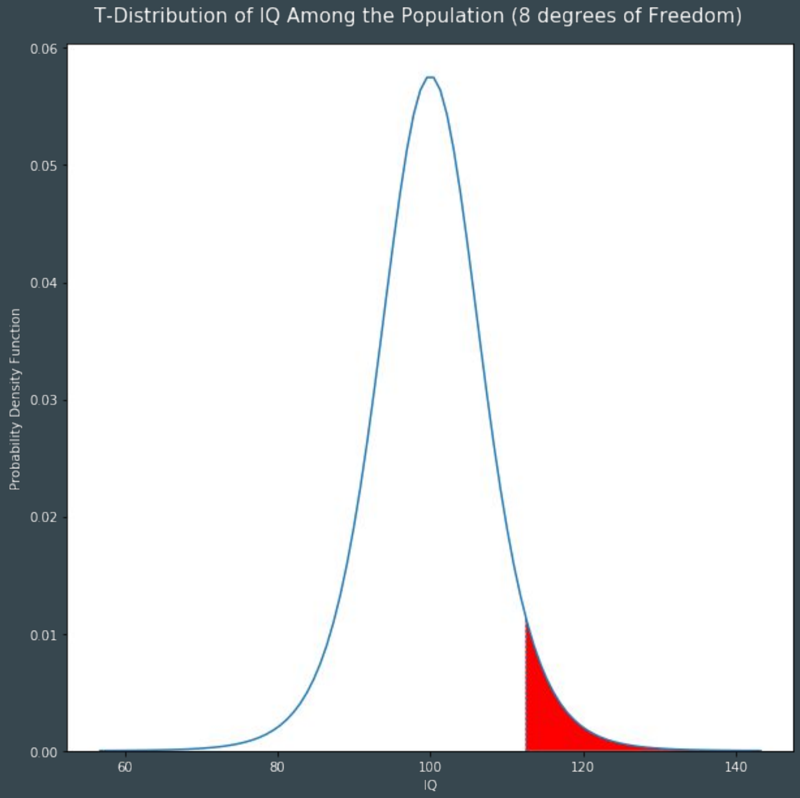


One-Sample T-Test

- We want to test if the mean IQ of the CCNY student is **greater than** the mean population IQ, given that IQ is normally distributed in the population with a hypothesized mean of 100 and an **unknown variance**.
- We take nine samples of CCNY students and find that they (you) have a mean IQ of 112 and a sample variance of 456 (and sample square root of 21.35)
- Given these parameters here - we want to do a one-sided test, but we do not know the inherent variance of the population, we can do a t-test with these parameters

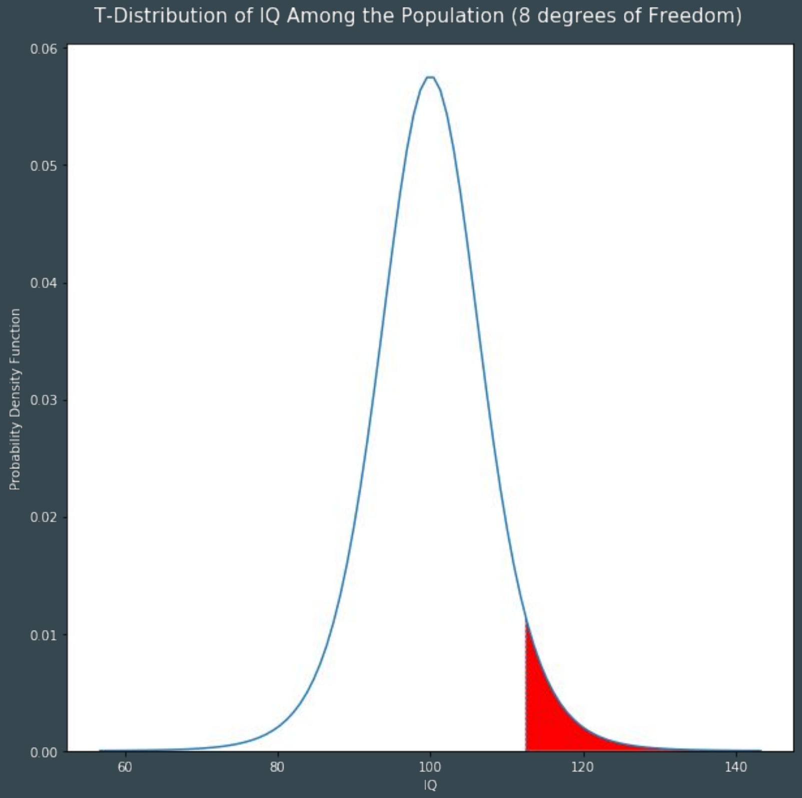
One Sample T-Test

- We can set up a t-distribution with a mean of 100, a standard deviation of 21.35 divided by the square root of our sample size (9) (around 7.11), and 8 degrees of freedom.
- The rejection region will be anything greater than the 95th percentile of this distribution, or 112.47.



One Sample T-Test

- Here, we will fail to reject the null hypothesis that the mean IQ of a CCNY student is equal to the mean IQ of the general population



Two-Sample T-Test

- What if we want to compare the means of two separate samples?
- We can do a **two-sample t-test** to do this.
- Let’s say we want to see if I am statistically running a faster mile in 2019 than I was in 2018
- I have ten samples from 2018 (in number of seconds):

470	444	476	511	441	441	513	481	431	472

- And ten samples from 2019 (in number of seconds):

420	394	426	461	391	391	463	431	381	422

Two-Sample T-Test

- The two sample t-test depends on a few assumptions:
  - Both samples are drawing from **normal distributions** with an unknown mean and variance
  - The variance of both samples are **the same**.
- The null hypothesis is that the differences between the two sample means equal a certain value
  - Often, that value is 0.
- Much like a one-sample test, the alternate hypothesis could be two-sided or one-sided to be greater or less.

Two-Sample T-Test

- I have ten samples from 2018 (in number of seconds):

420	394	426	461	391	391	463	431	381	422

- And ten samples from 2019 (in number of seconds):

381	381	410	323	331	378	359	413	364	344

- The first sample has a sample mean of 418
- The second sample has a sample mean of 368.4
- The difference between these means is -49.6
- The first sample has a sample variance of 836.67
- The second sample has a sample variance of 836.67

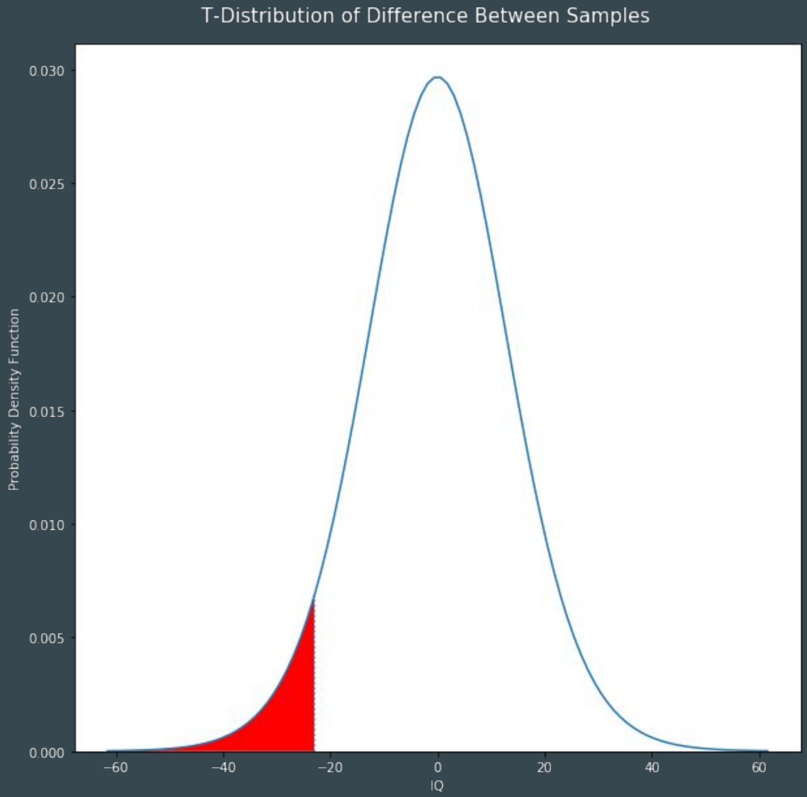
Two-Sample T-Test

- Because we are assuming that both samples have the same variance, we can come up with a *pooled variance* that takes into account both of our sample variances.
- It is not important to memorize this formula, but important to understand that we are using a single variance metric to account for both of our sample variances given the sample sizes of each sample
- With a pooled variance, we do not need to divide the sample variance by the length of the dataset like we would for a one-sample t-test
- Using this formula, our pooled variance will be 175.80

$$s_p^2 = \frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)$$

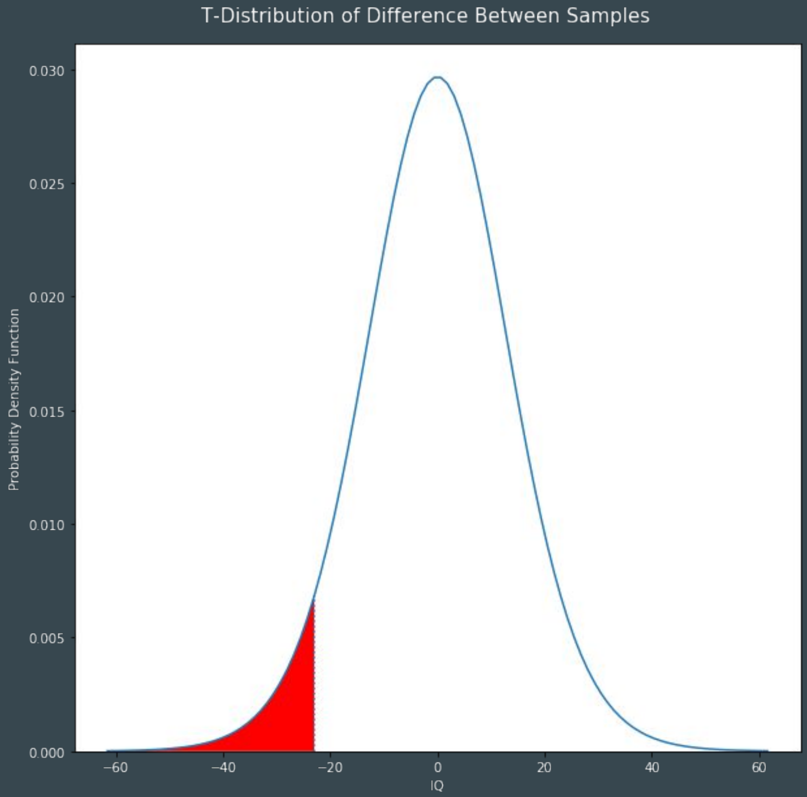
Two-Sample T-Test

- Now, we can build a T-distribution with a mean of 0, a standard deviation of 13.25, and 18 degrees of freedom (the length of our combined dataset minus two, since we are comparing two datasets)
- Now we can evaluate our test using a rejection region, just like a regular hypothesis test!



Two-Sample T-Test

- Specifically this is a one-tailed test, because our alternate hypothesis is that my running times in 2019 are lower than they were in 2018
- If the mean difference is anything less than -23 seconds, we can reject the null hypothesis that my mile time is the same in 2019 as it was in 2018
- Indeed, we can reject our null hypothesis since the mean difference was -50 seconds.



Bootstrap Sampling

- So far our hypothesis testing has been very strict about the assumptions of the original distribution
- The original distributions must be Normal, the variances must be equal, and we have only tested the difference in means
- **Bootstrap Testing** is a way of expanding what we can test via simulation, and will be useful for any hypothesis testing you do independently
- It is a new method that is remarkably simple, but only recently possible due to an increase in computing power

Bootstrap Sampling

- Bootstrap sampling takes the same premise as what we formally do with a hypothesis test. Given that the null hypothesis is true, what are the odds of our outcome happening?
- It does this, however, by physically simulating the process. In our case, we originally have two samples - 2018, which looks like this:

420	394	426	461	391	391	463	431	381	422

And 2019, which looks like this:

381	381	410	323	331	378	359	413	364	344

Bootstrap Sampling

- Bootstrap sampling takes the same premise as what we formally do with a hypothesis test. Given that the null hypothesis is true, what are the odds of our outcome happening?
- But now, we'll rearrange our samples, so that 2018 looks like this:

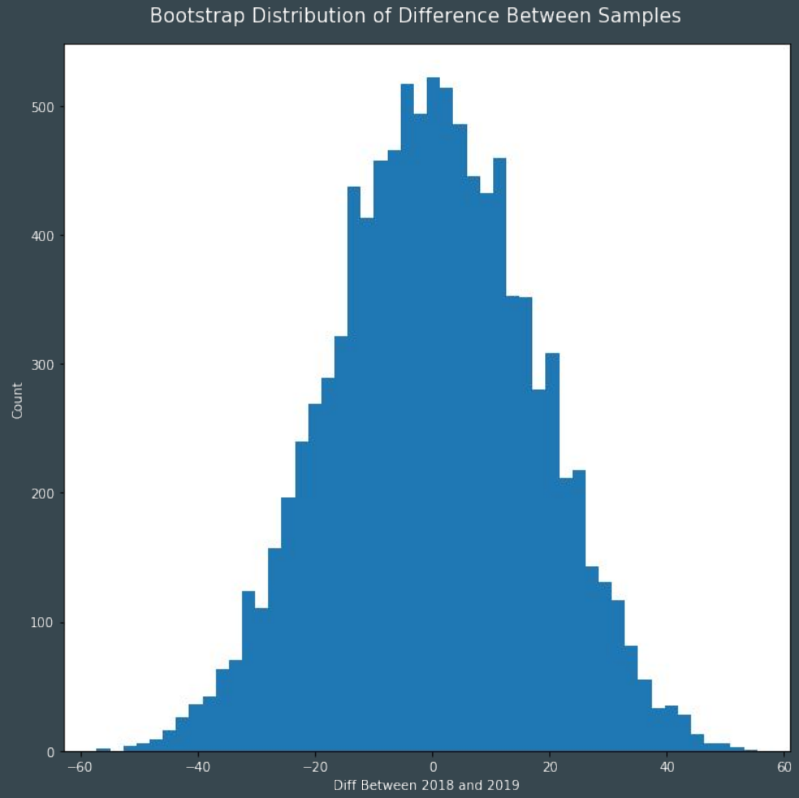
378	410	426	461	381	391	463	431	381	364

And 2019, which will now look like this:

391	381	394	323	331	420	359	413	422	391

Bootstrap Sampling

- We can do this 10,000 times, and see how many times the difference we saw a sample difference of -49.6 or less in our results.
- Specifically, there is a value of -49.6 or less only 8 times out of 10,000, equivalent to a p-value of 0.0008
- Note that we are seeing how many times this value exists in our simulations rather than modeling it as a percentile value of a distribution

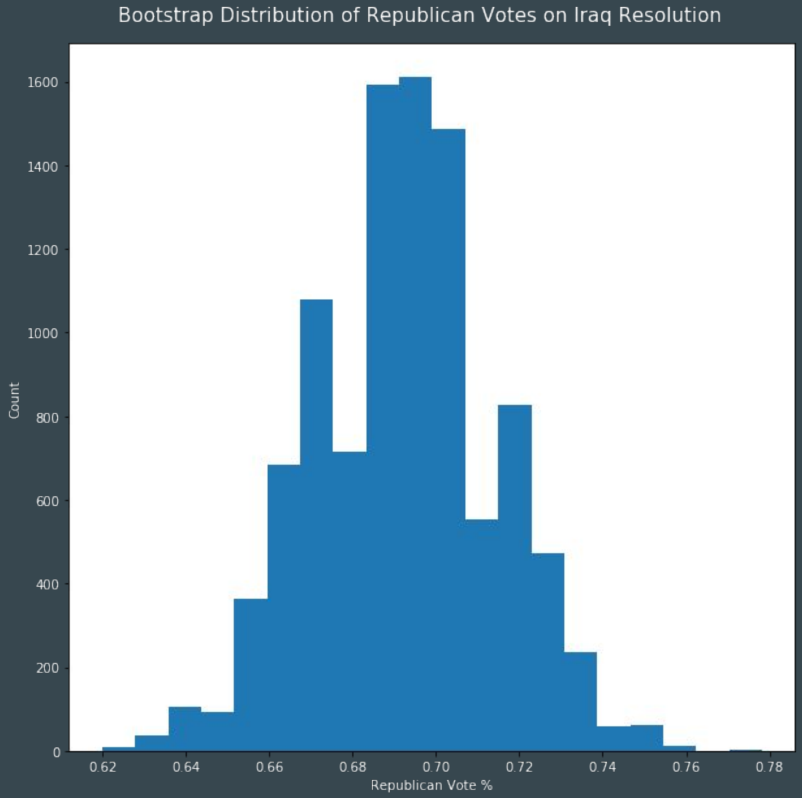


Bootstrap Sampling

- Let’s try a different type of problem. In 2002, Congress voted to go to war with Iraq in the ‘Authorization for Use of Military Force AGainst Iraq Resolution of 2002’.
- In the House of Representatives, Republicans voted 215-6 for the resolution, while Democrats voted 126-81 against the resolution.
- Did party affiliation make a difference in whether a representative voted for the resolution or not?
- We can do a bootstrap to find out, by shuffling Republicans and Democrats 10,000 times and seeing how many times “Republicans” vote at least 215 times.

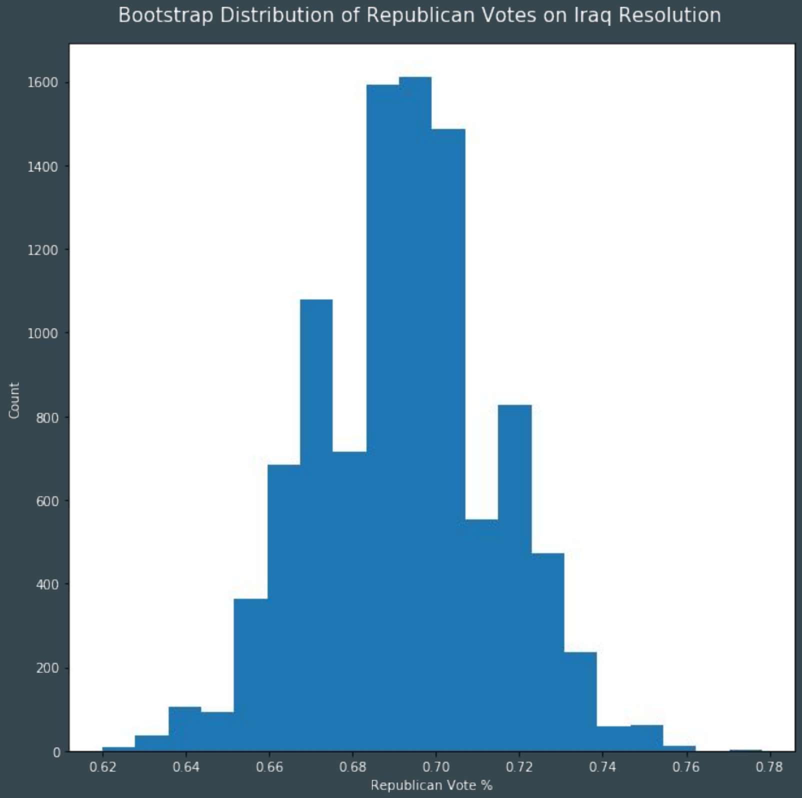
Bootstrap Sampling

- The null hypothesis is that party affiliation did not affect whether a representative would vote for the resolution
- The alternate hypothesis is that Republicans were *more likely* to vote for the resolution given their party affiliation.



Bootstrap Sampling

- After 10,000 simulations, we see that there are 215 or more Republican ‘Yes’ votes 0 times. It is statistically impossible that this happened randomly.
- The maximum number of Republican ‘Yes’ votes achieved through bootstrapping is 172, or 77.82%



Chi-Square Test for Goodness of Fit

- Last week, we looked at identifying rejection regions for a series of ten coin flips to determine whether the coin was fair or not
- We were able to manually create rejection regions because knew the PMF values for a binomial distribution with ten coin flips and a 50% probability of success on each flip

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

Chi-Square Test for Goodness of Fit

- What if we didn't know the PMFs? We can do a **chi-square test for the goodness of fit** that evaluates the observed results of a discrete outcome versus the expected results.
- Specifically, we calculate the sum of the squared difference between each observed and expected value divided by the expected value. This is called **Pearson's Chi-Square statistic**.
- We then see if that value fits the rejection region for a **chi-square distribution** with the given number of degrees of freedom

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Square Test for Goodness of Fit

- Here, the null hypothesis is that the observed values come from the listed distribution.
- The alternate hypothesis is that the observed values come from another distribution.
- If it’s in the rejection region, we **reject the null hypothesis** that the observed values come from the listed distribution.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Square Test for Goodness of Fit

- Say I roll a pair of dice 100 times. How will I know if it is fair or unfair?
- If my results are:

Value	2	3	4	5	6	7	8	9	10	11	12
Count	7	10	9	9	10	6	9	14	10	9	7

- And the expected results are:

Value	2	3	4	5	6	7	8	9	10	11	12
Count	2.7	5.5	8.3	11.1	13.8	16.6	13.8	11.1	8.3	5.5	2.7

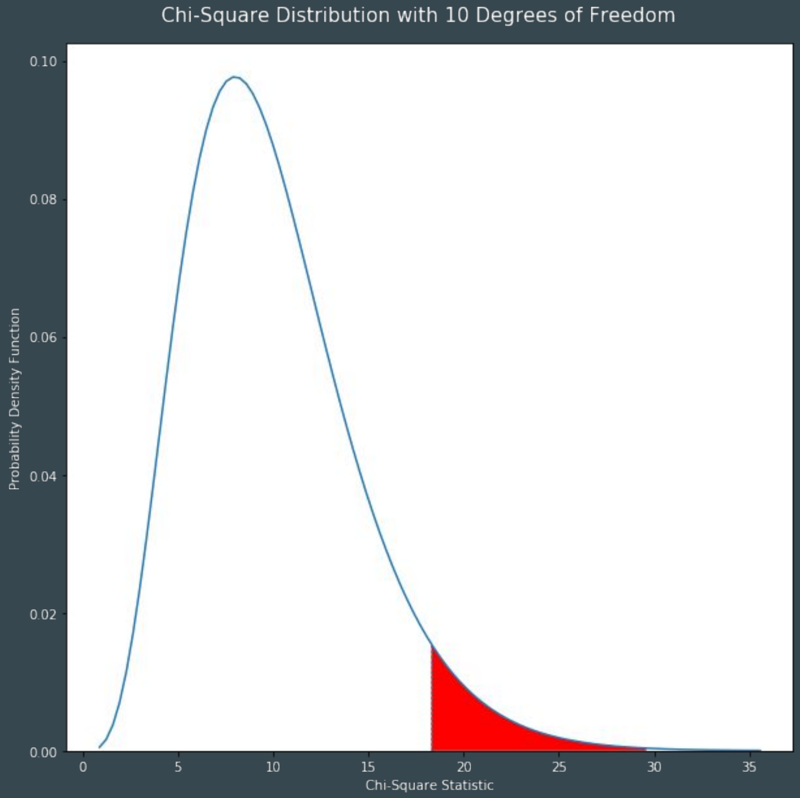
Chi-Square Test for Goodness of Fit

- Can we fail to reject the null hypothesis that our results come from two fair die?
- Pearson’s chi-square statistic for this is 29.702.
- We can then see whether this value falls in the **rejection region** for a Chi-Square distribution with **ten** degrees of freedom.

$$\frac{(7-2.7)^2}{2.7}+\frac{(10-5.5)^2}{5.5}+\frac{(9-8.3)^2}{8.3}+\ldots+\frac{(7-2.7)^2}{2.7}$$

Chi-Square Test for Goodness of Fit

- With a p-value of less than 0.001, we can **reject the null hypothesis** that our two dice are fair.
- To the right we see that the rejection region of the 95th percentile or higher incorporates any value that's roughly ~19 or higher.



Chi-Square Test for Homogeneity

- Another chi-square test involves testing multiple independent different data sets to see if they are drawn from the same discrete distribution
- Here, the null hypothesis is that each dataset is drawn from the same distribution
- The alternate hypothesis is that each dataset is not drawn from the same distribution.
- Essentially, we will find the expected values for the two datasets, assuming they're drawn from the same distribution, and do a chi-squared test on it, similar to how we did for the goodness of fit test.
- The chi-square test will have (n - 1) * (m-1) degrees of freedom, where n is the number of observations in each distribution, and m is the number of distributions.

Chi-Square Test for Homogeneity

- Say that someone finds to find a long-lost work by Shakespeare.
- We want to try and verify if this work is by Shakespeare by comparing the frequency of common words to see if the relative frequencies are similar to what Shakespeare would have written.

Word	*a*	*an*	*this*	*that*
King Lear	150	30	30	90
Lost Work	90	20	10	80

Chi-Square Test for Homogeneity

- Say that someone finds to find a long-lost work by Shakespeare.
- We want to try and verify if this work is by Shakespeare by comparing the frequency of common words to see if the relative frequencies are similar to what Shakespeare would have written.

Word	*a*	*an*	*this*	*that*	Total
King Lear	150	30	30	90	300
Lost Work	90	20	10	80	200
Total	240	50	40	170	500

Chi-Square Test for Homogeneity

- Say that someone finds to find a long-lost work by Shakespeare.
- We want to try and verify if this work is by Shakespeare by comparing the frequency of common words to see if the relative frequencies are similar to what Shakespeare would have written.

Word	*a*	*an*	*this*	*that*	Total
King Lear	150 (144)	30 (30)	30 (24)	90 (102)	300
Lost Work	90 (96)	20 (20)	10 (16)	80 (68)	200
Total	240	50	40	170	500

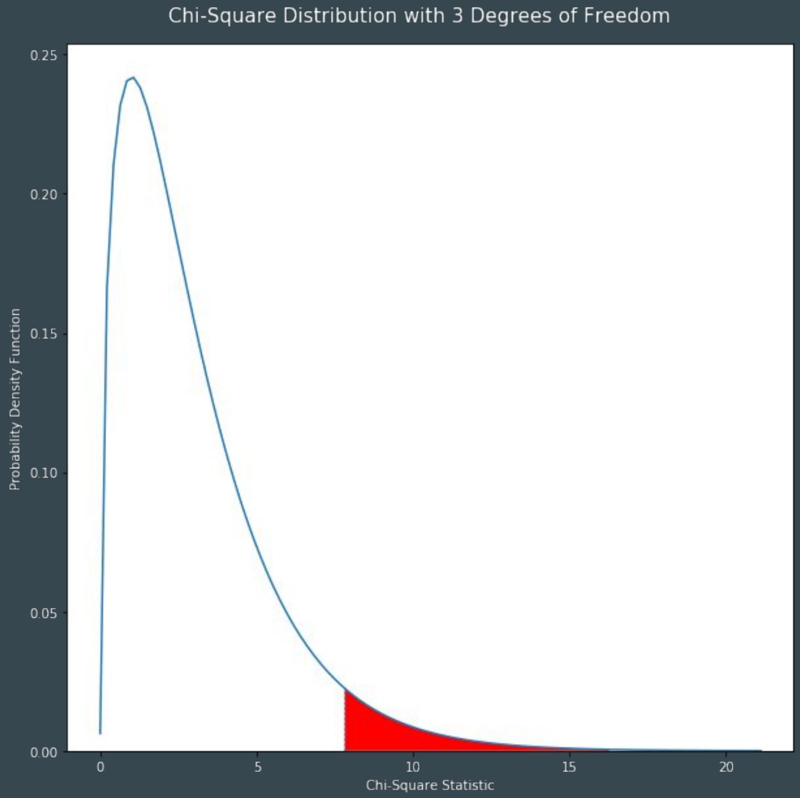
Chi-Square Test for Homogeneity

- The Pearson's chi-square statistic for this is 7.90
- Since we are comparing two sets of four observations, we will have (4 - 1) * (2-1), or 3, degrees of freedom.
- We can then see whether this value falls in the **rejection region** for a Chi-Square distribution with **three** degrees of freedom.
- The null hypothesis is that both works come from the same distribution, the alternate hypothesis is that they do not.

$$\frac{(150-144)^2}{144}+\frac{(90-96)^2}{96}+\frac{(30-30)^2}{30}+\frac{(20-20)^2}{20}+\frac{(30-24)^2}{24}+\frac{(10-16)^2}{16}+\frac{(90-102)^2}{102}+\frac{(80-68)^2}{68}$$

Chi-Square Test for Goodness of Fit

- With a p-value of less than 0.048, we can **reject the null hypothesis** that both works come from the same distribution, or in this case that they are written by the same author (Shakespeare).
- To the right we see that the rejection region of the 95th percentile or higher incorporates any value that's roughly ~8 or higher.


Chi-Square Distribution with 3 Degrees of Freedom
Probability Density Function
Chi-Square Statistic