# Inferential Statistics Report

Minerva Schools at KGI

CS50: Formal Analyses

Prof. P. Watson

December 16, 2020

## 1. Introduction

In my high school senior year, my parents and I talked about the universities that best suit me for the future. After graduation, the median income cannot be disregarded since wealth is one of the factors that can alter our life. This report explores the difference of Starting Median Salary (SMS) in two different regions in the United States: Northeastern and Southern. I chose these two regions since I have lived in the Northeastern area for a long time while I have not visited the Southern region. Exploring between known and unknown regions will give me a better understanding of how the regions are different in terms of people's median salary.

The dataset I chose includes different schools from regions in the United States and their starting median salary, mid-career median salary, mid-career 10th Percentile salary, mid-career 25th Percentile salary, mid-career 75th Percentile salary, and mid-career 90th Percentile salary. This paper researches how Northeastern and Southern Universities share the difference in their starting median salary.

The null hypothesis is that there is no difference in the mean starting median salary between Northeastern and Southern Universities. The alternative hypothesis is there is a difference in mean starting median salary between Northeastern and Southern Universities.

- H0: Northeastern - Southern = 0
- HA: Northeastern - Southern != 0

## 2. Dataset

The dataset is adapted from Kaggle. The assignment instruction provided the CSV data file, and I used this in this research. Between the files, I chose 'Salaries By Region' to filter the data only for the Northeastern and Southern regions to take a closer look at my research question. The research question is: "is there any significant evidence that shows the critical difference between the mean of starting median salary for the universities in Northeastern and Southern regions?"

The regions are the qualitative nominal variables because the Northeastern and Southern regions only provide each region's labeling but not any numerical values. The starting median salaries, in general, are continuous quantitative variables. However, for this data set, one can realize that median salaries are rounded up to hundreds of dollars. This means that this variable's step increases by one hundred dollars, so the starting median salary is a quantitative discrete variable.

According to the National Association of Colleges and Employers, students who graduated with STEM degrees generally have a greater average salary than students who graduated with non-STEM degrees (National Association of Colleges and Employers, 2019). Therefore, the percentage of STEM students from each college can be a confounding variable. If college A offers STEM programs in general, they are more likely to have a bigger median starting salary than college B that does not offer many STEM programs. However, in this paper, one is only interested in how the college's region affects students' starting salary regardless of their major; therefore, the paper assumes the starting median salary as a whole university and its region. The imported data can be viewed in Appendix A.[1]

## 3. Analysis

### Summary Statistics

I used Python to read the .csv data file. The data was filtered already with the Northeastern and Southern schools only. Table 1 summarizes the key statistics for the starting median salary for Northeastern and Southern university graduates.

| **Table 1: Quantities of Northeastern and Southern Universities in Data** | |
|---|---|
| Northeastern | Southern |
| 82 | 71 |
| 53.6% (82/(82+71))*100 | 46.4% (71/(82+71))*100 |

[1] **#variables:** Identified and clarified the different types of variables. Explained why those variables affect this research and significance. Included the possible confounding variable and addressed it once again in results so that how this confounding variable can be handled further to conduct the more accurate research.

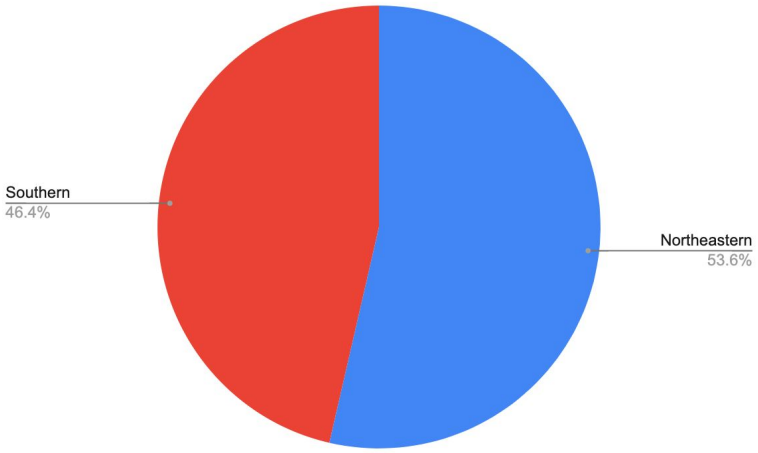



**Figure 1:** Pie Chart for the percentage of each region in the data set.

| **Table 2: Summary Statistics for the variables of interest** | |
|---|---|
| | Starting Median Salary for Northeastern/Southern University Students ($ per year) |
| Count | $n = 153$ |
| Mean | $\bar{x}_{salary} = 46947.06$ |
| Median | 45600 |
| Mode | 36900, 41100, 42100, 42800, 43100, 45400, 45700, 48000 |
| Standard Deviation | $s_{salary} = 6847.3$ |
| Range | 72200 - 34800 = 37400 |

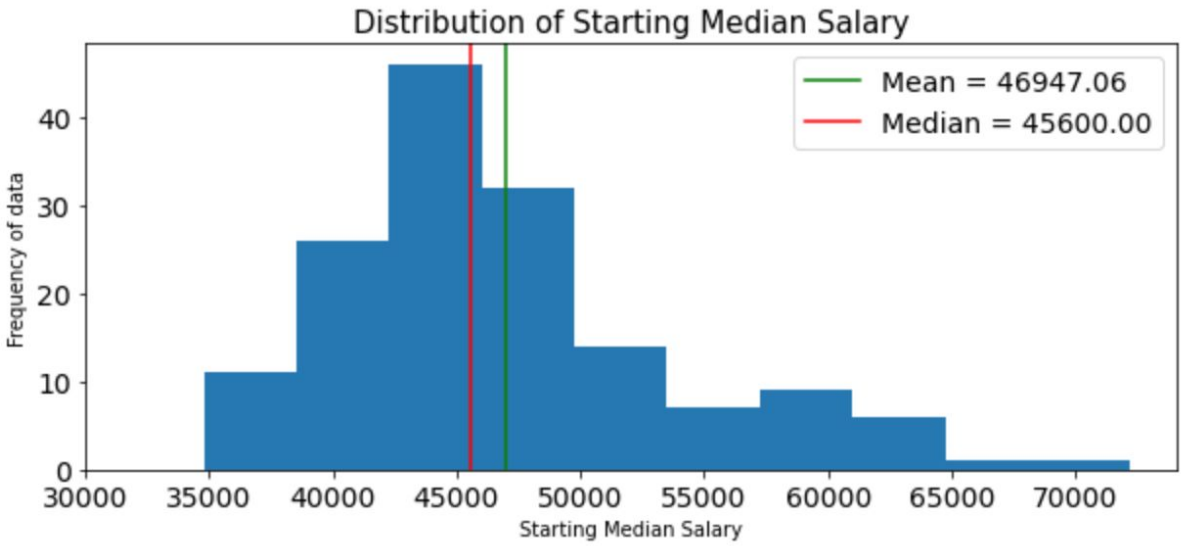



**Figure 2:** Histogram for both Northeastern/Southern Universities

Based on the histogram, one can infer that the histogram is right-skewed. This means that most students earn equal to or less than mean salaries as the median is lower than the mean value. The range is $37,400, which means the gap between the school has the maximum starting median salary and the school has the minimum starting median salary is huge. It is most likely due to the data set, including the schools with known-name values, including some colleges that are not popular. Table 2 and Figure 2's python code can be viewed in Appendix B and C.[2][3]

### Confidence Interval

Since the current dataset does not include every university in the Northeastern and Southern area, the current data is a sample. To estimate the mean starting median salary for the population (every university in both regions), I compute a 95% confidence interval. This will give a plausible range of values. This confidence interval will give me 95% confidence in the mean value of the population.

Confidence interval can be calculated as:

$$\overline{x} \pm z \frac{s}{\sqrt{n}}$$

[2] **#descriptivestats:** Interpreted the descriptive statistic values. Accurately calculated those values with python. Explained the significance of them.
[3] **#dataviz:** Effectively generated the histogram to summarize the data set. Identified the mean and median value to describe the histogram.

$z\frac{s}{\sqrt{n}}$ in the formula is a standard error (SE). $\bar{x}$ is sample mean, z is z-score, s is standard deviation for the sample, and n is the size of the sample. However, in this case, I use t-value instead of z score since I don't know the standard deviation for the population.

To proceed, I have to make sure that the samples are randomly chosen and distributed normally with enough sample size (n>30). Even if we have enough sample size, the data sets are not necessarily chosen randomly and right-skewed. Technically, I cannot proceed with this but assume the conditions are met in this report. Therefore, the confidence interval and the significance level here in this report assume that the distribution is normal, and the sample is chosen randomly.[4]

The confidence interval here is (45849.78, 48044.34). It means there is 95% confidence that the population means are located between those two values. Python code for calculating this confidence interval can be found in Appendix F.[5]

**Table 3: Summary Statistics for Each Region**

| | Starting Median Salary for Northeastern University Students ($ per year) | Starting Median Salary for Southern University Students ($ per year) |
|---|---|---|
| Count | $n_1 = 82$ | $n_2 = 71$ |
| Mean | $\bar{x}_1 = 48679.27$ | $\bar{x}_2 = 44946.48$ |
| Median | 46700 | 44100 |
| Mode | 45700, 48000 | 43100 |
| Standard Deviation | $s_1 = 7495.2$ | $s_2 = 5355.89$ |
| Range | 72200 - 36900 = 35300 | 64000 - 34800 = 29200 |

[4] **#distributions:** Explained the data is right-skewed. However, stated the assumption that I count this data as normal distribution to conduct further calculations.

[5] **#confidenceintervals:** Accurately interpreted and calculated the confidence interval using python. Explained the meaning of the confidence intervals and its assumption.

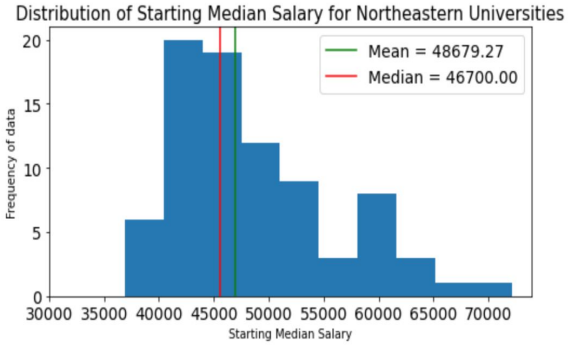



**Figure 3: Histogram for Northeastern Universities' Starting Median Salary**

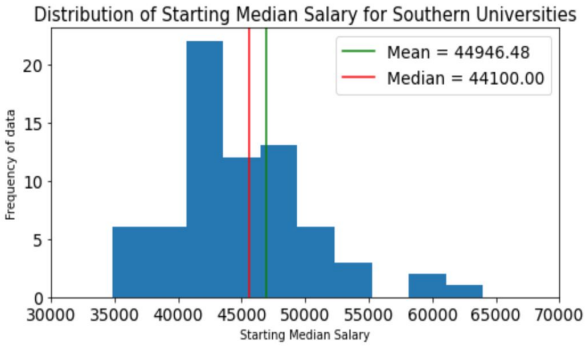



**Figure 4: Histogram for Southern Universities' Starting Median Salary**

Both histograms are right-skewed. While some northeastern colleges have a $60000s+ starting median salary, only a few southern schools have a $60000s+ starting median salary. This means the southern schools with a high starting salary could be outliers, while most southern schools have relatively lower starting median salary. With this, we can infer that Southern universities generally have lower starting median salaries than Northeastern universities. Python code for Table 3, Figure 3, and 4 can be found in Appendix D and E.[6][7]

### Confidence Interval

With the same assumption as stated above, I compute a confidence interval for each Northeastern and Southern Universities. This will give a plausible range of the values for the population mean value. I set our confidence level to 95%.

Both data have enough sample size: more than 30.

The confidence interval for the northeastern universities is (47022.26, 50336.28). The confidence interval for the southern universities is (43669.74, 46223.22). It means there is 95% confidence that the population means is located between those two values.

---

[6] **#descriptivestats:** Interpreted the descriptive statistic values. Accurately calculated those values with python. Explained the significance of them.

[7] **#dataviz:** Effectively generated each histogram to summarize the data set. Compared the data visualization to come up with a possible hypothesis. Identified the mean and median value to describe the histogram.

Python code of this confidence interval can be found in Appendix F.[8]

**The difference of Means Test**

To answer whether there is a statistically significant difference in starting median salaries between northeastern and southern universities, I have to perform a difference of means significance test. Regions are split based on the geographical factors: northeastern and southern. I set my significance level: $\alpha = 0.05$ .

Hypotheses are clearly set once again:

- H0: Northeastern Universities Mean Starting Median Salary - Southern Universities Mean Starting Median Salary = 0
- HA: Northeastern Universities Mean Starting Median Salary - Southern Universities Mean Starting Median Salary != 0

Since I am looking for the difference in both ways, I use a two-tailed test.

I have enough sample sizes here ( $n_1, n_2 > 30$ ). Since I do not know the standard deviations for the population, I use t distribution for inference. Although big enough sample sizes will not significantly differ between using z distribution and t distribution, it is still better to use t distribution. If I use z distribution, I have to estimate the population's standard deviation. In this estimation process, I expect some errors can occur; therefore, I avoid using z distribution but use t distribution.

In order to proceed with the t-score, I use the formula: $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ . After the calculation, I get t = 3.55.

Because my hypothesis considers differences in either direction, I conduct the two-tails test. After conducting a statistical significance test and the differences of the mean test, the p-value is calculated as 0.000685. Since our significance level is 0.05, and the p-value is lower than the significance level, I reject the null hypothesis.

---

[8] **#confidenceintervals:** Accurately interpreted and calculated the confidence interval using python. Explained the meaning of the confidence intervals and its assumption. Stated the differences between the confidence intervals in each region.

Mentioned above is a statistical significance. Since I am looking at the real-world significance, it is important to understand the practical significance in this report. Since my null hypothesis is rejected, I may do a practical significance test to examine how the mean of starting median salaries is affected. I calculated the effect size with python; Cohen's d value resulted in 0.563. Even though my sample size is big enough, I conduct the Hedge's g calculation slightly more accurately. The Hedge's g is resulting in 0.557. I understand that this effect size tells us that there is a moderate difference between the mean starting median salaries between northeastern universities and southern universities. The test can be found in Appendix G.[9]

## 4. Results and Conclusions

From seeing the difference between the 95% confidence interval between the Northeastern and Southern Universities Starting Median Salary, one can think that there could be an actual difference between the salaries in general. From computing the differences of the mean test, I observed that our p-value is way smaller than our significance level; therefore, I conclude that there is a difference. By applying Cohen's D and Hedge's G, I interpreted how it differs in our real-world. Since the Hedge's g resulted in 0.557, one can conclude a moderate difference in starting median salaries between the Northeastern and Southern Universities.

The research showed that Northeastern Universities have a greater starting median salary than Southern Universities. This is a strong inductive argument because it provided the statistical calculations with relevant variables. It successfully analyzed the population with sampled data, which means there is high confidence that the whole population was represented in this paper. If the sample size was too small, it would be not good inductive reasoning; however, since we have more than enough sample size, this is good reasoning because the bigger sample size represents the whole population better.

This research concludes that there is a moderate difference in starting median salaries between the Northeastern Universities and the Southern Universities. The Northeastern Universities' possible reason for having a higher starting median salary is some well-known schools like Ivy-League schools, and other top national universities are mostly located in the Northeastern region. It affected the nearby universities because universities near each other co-work to make their students' future better. Although there are a few universities with more than $60,000+ starting median salaries in the Southern area; however, they are almost outliers because there are

[9] **#significance:** By calculating the differences of the means test and stating Cohen's D and Hedge's G, accurately interpreted the differences. Stated the moderate difference between the statistical significance and practical significance.

very few. Considering this, the Southern Universities, in general, have an even lower starting median salary than the Northeastern Universities.

This research can be further developed if the data states the students' majors' starting median salary. It is better to compare with the same major since that can be a control variable that makes the research stronger.[10]

## 5. References

Diez, D., Barr, C., & Cetinkaya-Rundel, M. (2015). *OpenIntro Statistics* (3rd ed.).

Kaggle. (2017, April 29). *Where it Pays to Attend College*. https://www.kaggle.com/wsj/college-salaries

National Association of Colleges and Employers. (2019, January 9). *STEM Majors Projected to Be Class of 2019's Top Paid*. https://www.naceweb.org/job-market/compensation/stem-majors-projected-to-be-class-of-2019s-top-paid/

## 6. Reflection

I received feedback that my inductive reasoning is strong but failed to show how the premises are highly correlated with the conclusion. Therefore, I focused on providing strong and sound premises based on the statistical data that I observed and connected to the real world in order to show how the statistics actually represent our reality.

Before applying #induction, I reviewed the class that we were introduced to #induction. It was a good step to review my knowledge by applying #scienceoflearning.

I had a hard time at the end of this semester because of my personal feeling. I could not get over this emotional loop, and it definitely made me unproductive. As a result, I got accommodation to

---

[10] **#induction:** Based on the calculation, I came up with strong premises and resulted in a strong conclusion. The inductive reasoning here is strong since they are all based in statistical evidence.

extend my deadline, but I still could not use this time efficiently. I was unhappy with relatively lower effort compared to my previous assignment.

This is totally my fault that I could not spare hours for this assignment even though this assignment was released a while ago. If I planned at least a few hours from weeks ago, I think I could finish this assignment without accommodation and with high quality.

With this experience, I learned that sparing my time when I am emotional is especially important because this is the only way that I can get over it.[11][12][13][14][15]

[11] **#selfawareness:** Reviewed the class and feedback to come up with a better outcome for this assignment. Stated the initial circumstances in my life and reflected how that affected my assignment.
[12] **#professionalism:** Worked on the sample assignment to meet all the formatting requirements. Properly cited the sources that I used outside of the class. Proofread by the SSS tutor. Used Grammarly to make sure there are no grammar issues.
[13] **#audience:** Understand that this assignment can be viewed by non-experts, properly explain the all mathematical interpretation in human language so that the audience can easily follow.
[14] **#composition:** Used the clear sentences and easy words so that the ones who are not familiar with statistics can easily understand and follow.
[15] **#organization:** Used the report organization so that the audience can understand step-by-step. From introduction to the conclusion, the report supports the claim made in each section.

# 7. Appendix

## Base

The full Jupyter notebook file and other data can be accessed here.

## Appendix A: Import and Analyze Data

In [125]:
```
#import all the necessary libraries

%matplotlib inline
import pandas as pd
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
plt.rcParams.update({'font.size': 14})

#import the data set
data = pd.read_csv("college data set - Sheet2.csv", sep=',')
salary = data["Starting Median Salary"]
data.head(7)
```

Out[125]:

| | Region | Starting Median Salary |
|---|---|---|
| 0 | Southern | 64000 |
| 1 | Southern | 55000 |
| 2 | Southern | 58900 |
| 3 | Southern | 58300 |
| 4 | Southern | 51200 |
| 5 | Southern | 52700 |
| 6 | Southern | 47000 |

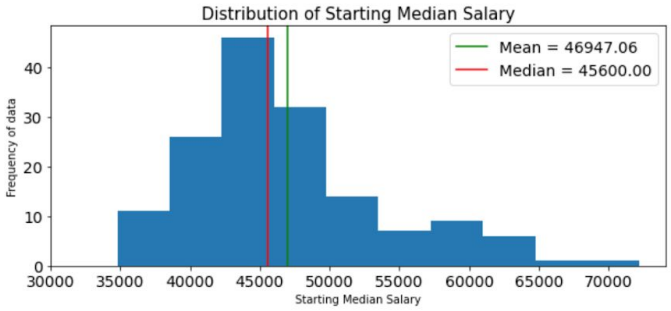
## Appendix B: Descriptive Statistics as a Whole

In [126]:
```
salary_mean = round(salary.mean(),2) #round up to two digits
salary_median = salary.median()
sd_salary = round(np.std(salary),2) #round up to two digits

def mode(lst): #since I have multiple modes, I created another function for mode
    L1=[] #count the occurrence of each number and append or add findings to the list

    i = 0 #count the numbers and put them into L1
    while i < len(lst) :
        L1.append(lst.count(lst[i]))
        i += 1

    # the occurrences for each number in sorted lst
    # create a custom dictionary d1 for k : V
    # k = value, v = occurence

    d1 = dict(zip(lst, L1))
    d2=[k for (k,v) in d1.items() if v == max(L1)] # the k values with the highest v values.
    return d2

print("Your # of data is:",len(salary))
print("Mean is:", salary_mean)
print("Median is:", salary_median)
print("Mode is:", sorted(mode(salary.to_list())))
print("Standard Deviation is:", sd_salary)
print("Your range is:", max(salary) - min(salary))
```

```
Your # of data is: 153
Mean is: 46947.06
Median is: 45600.0
Mode is: [36900, 41100, 42100, 42800, 43100, 45400, 45700, 48000]
Standard Deviation is: 6847.3
Your range is: 37400
```

# Appendix C: Data Visualization as a Whole

In [127]:
```
plt.hist(salary, alpha=1, linewidth=5)
plt.title("Distribution of Starting Median Salary", fontsize = 15)
plt.xlabel("Starting Median Salary", fontsize = 10)
plt.ylabel("Frequency of data", fontsize = 10)
plt.gcf().set_figwidth(10)

plt.xticks([30000,35000,40000,45000,50000,55000,60000,65000,70000]) #increase xticks by 10000
plt.axvline(salary_mean, color='g', label = "Mean = 46947.06") #label legend for mean
plt.axvline(salary_median, color='r', label = "Median = 45600.00")#label legend for median
plt.legend()
ax=plt.gca()
ax.set_facecolor('w')
plt.show()
```



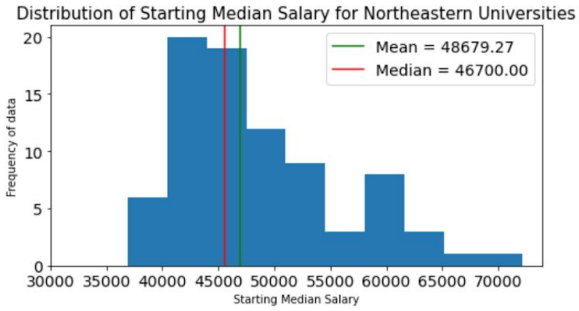

# Appendix D: Descriptive Statistics for Each Region

In [128]:
```
dataN = pd.read_csv("college data set - Northeastern.csv", sep=',')
salaryN = dataN["Starting Median Salary"]

salaryN_mean = round(salaryN.mean(),2) #round up to two digits
salaryN_median = salaryN.median()
sd_salaryN = round(np.std(salaryN),2) #round up to two digits

def mode(lst): #since I have multiple modes, I created another function for mode
    L1=[] #count the occurrence of each number and append or add findings to the list

    i = 0 #count the numbers and put them into L1
    while i < len(lst) :
        L1.append(lst.count(lst[i]))
        i += 1

    # the occurrences for each number in sorted lst
    # create a custom dictionary d1 for k : V
    # k = value, v = occurence

    d1 = dict(zip(lst, L1))
    d2=[k for (k,v) in d1.items() if v == max(L1)] # the k values with the highest v values.
    return d2

print("Your # of data is:",len(salaryN))
print("Mean is:", salaryN_mean)
print("Median is:", salaryN_median)
print("Mode is:", sorted(mode(salaryN.to_list())))
print("Standard Deviation is:", sd_salaryN)
print("Your range is:", max(salaryN) - min(salaryN))
```

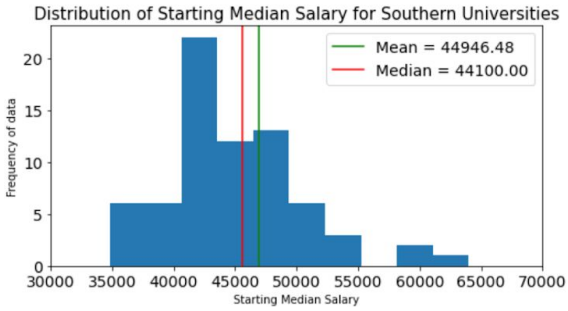
In [130]:
```
dataS = pd.read_csv("college data set - Southern.csv", sep=',')
salaryS = dataS["Starting Median Salary"]

salaryS_mean = round(salaryS.mean(),2) #round up to two digits
salaryS_median = salaryS.median()
sd_salaryS = round(np.std(salaryS),2) #round up to two digits

def mode(lst): #since I have multiple modes, I created another function for mode
    L1=[] #count the occurrence of each number and append or add findings to the list

    i = 0 #count the numbers and put them into L1
    while i < len(lst) :
        L1.append(lst.count(lst[i]))
        i += 1

    # the occurrences for each number in sorted lst
    # create a custom dictionary d1 for k : V
    # k = value, v = occurence

    d1 = dict(zip(lst, L1))
    d2=[k for (k,v) in d1.items() if v == max(L1)] # the k values with the highest v values.
    return d2

print("Your # of data is:",len(salaryS))
print("Mean is:", salaryS_mean)
print("Median is:", salaryS_median)
print("Mode is:", sorted(mode(salaryS.to_list())))
print("Standard Deviation is:", sd_salaryS)
print("Your range is:", max(salaryS) - min(salaryS))
```

# Appendix E: Data Visualization for Each Region

In [129]:
```
plt.hist(salaryN, alpha=1, linewidth=5)
plt.title("Distribution of Starting Median Salary for Northeastern Universities", fontsize = 15)
plt.xlabel("Starting Median Salary", fontsize = 10)
plt.ylabel("Frequency of data", fontsize = 10)
plt.gcf().set_figwidth(8)

plt.xticks([30000,35000,40000,45000,50000,55000,60000,65000,70000]) #increase xticks by 10000
plt.axvline(salary_mean, color='g', label = "Mean = 48679.27") #label legend for mean
plt.axvline(salary_median, color='r', label = "Median = 46700.00")#label legend for median
plt.legend()
ax=plt.gca()
ax.set_facecolor('w')
plt.show()
```



In [131]:
```
plt.hist(salaryS, alpha=1, linewidth=5)
plt.title("Distribution of Starting Median Salary for Southern Universities", fontsize = 15)
plt.xlabel("Starting Median Salary", fontsize = 10)
plt.ylabel("Frequency of data", fontsize = 10)
plt.gcf().set_figwidth(8)

plt.xticks([30000,35000,40000,45000,50000,55000,60000,65000,70000]) #increase xticks by 10000
plt.axvline(salary_mean, color='g', label = "Mean = 44946.48") #label legend for mean
plt.axvline(salary_median, color='r', label = "Median = 44100.00")#label legend for median
plt.legend()
ax=plt.gca()
ax.set_facecolor('w')
plt.show()
```




# Appendix F: Confidence Interval

In [132]:
```
#confidence interval 1

stats.t.interval(alpha=0.95, df=len(salary)-1, loc=np.mean(salary), scale=stats.sem(salary))
```

Out[132]: (45849.77964777234, 48044.33799928649)

In [133]:
```
#confidence interval northeastern

stats.t.interval(alpha=0.95, df=len(salaryN)-1, loc=np.mean(salaryN), scale=stats.sem(salaryN))
```

Out[133]: (47022.25808619383, 50336.278499172025)

In [134]:
```
#confidence interval southern

stats.t.interval(alpha=0.95, df=len(salaryS)-1, loc=np.mean(salaryS), scale=stats.sem(salaryS))
```

Out[134]: (43669.73673807361, 46223.221008405264)

# Appendix G: Differences of the Means Test

In [135]:
```
def dif_of_means_test(data1,data2,tails):
    n1 = len(data1) #length of the data is equal to the sample size
    n2 = len(data2)

    x1 = np.mean(data1) #use numpy to get mean here
    x2 = np.mean(data2)

    s1 = np.std(data1,ddof=1) # Having Bessel's correction here
    s2 = np.std(data2,ddof=1) # Therefore I use (n-1) as my denominator

    SE = np.sqrt(s1**2/n1 + s2**2/n2)
    Tscore = np.abs((x2 - x1))/SE
    df = min(n1,n2) - 1 # Using conservative estimation as Open Intro said.
    pvalue = tails*stats.t.cdf(-Tscore,df)

    SDupdated = np.sqrt((s1**2*(n1-1) + s2**2*(n2-1))/(n1+n2-2)) # OpenIntro section 5.3.6
    Cohensd = (x2 - x1)/SDupdated
    HedgesG = Cohensd*(1-(3/((4*df)-1)))

    print('t =', Tscore)
    print('p =', pvalue)
    print('d =', np.abs(Cohensd)) #get an absolute value
    print('g =', np.abs(HedgesG))


dif_of_means_test(salaryN,salaryS,2) #have two tails
```

```
t = 3.5536687864510905
p = 0.0006852898454314402
d = 0.5627700377955126
g = 0.5567187470665286
```

---

# Resources

- An example for CS50
    - Used the example guide to follow the formatting and the writing. With this guide, I got an idea for how to cite the sources and put the diagrams effectively.