Statistical Inference

Minerva Schools at KGI

CS50

Prof. Stan

December 10, 2020

Statistical Inference

Introduction

The paper explores Where it Pays to Attend College dataset to evaluate how the salaries differ in two regions of the US (The Wall Street Journal, 2020). The primary samples are Mid-Career 90th Percentile Salaries in Northeastern and Southern US. The entries were used to investigate whether the difference in means of Mid-Career 90th Percentile Salaries in these regions is statistically significant to conclude that the mid-career salaries are generally higher in Northeastern. First, the confidence interval for a difference in means of median salaries was obtained. After checking the conditions and evaluating assumptions, tests for statistical (t-test) and practical (Cohen's d, Hedge's g) significance were performed. It was necessary to determine the extent to which universities' location influences the distribution of salaries across the US. In other words, we seek the answer to the question "Is there convincing evidence that graduates from universities in Northeastern earn more than university students in the Southern states?" Ultimately, this analysis will perform the mentioned statistical calculations to check the null hypothesis and provide more insight into the mid-career salaries in the US. As a result, the following hypothesizes were stated.

Null: There is no statistically significant evidence that the mean of starting median salary for Northeastern schools is larger than that of Southern schools.

$$\mu_{northeastern} - \mu_{southern} = 0$$

Alternative: There is statistically significant evidence that the mean of starting median salary for Northeastern schools is larger than that of Southern schools.

$$\mu_{northeastern} - \mu_{southern} > 0$$

Dataset

The data is provided by The Wall Street Journal and includes 273 universities from four regions, including California, South, Northeast, West, Midwest, and graduates' salaries (starting Median, Mid-career Median, Mid-career 10th, 25th, 75th, and 90th percentiles) from these regions. The research question is: "Is there any statistically significant evidence that the mean of starting median salary for Midwestern schools is larger than that of Southern schools?" The dataset is complete and does not need any pre-processing or cleaning.

The sample was used to investigate whether the mid-career 90th percentile in the Northeast is higher than in the South. The region is a qualitative nominal variable since there is no intrinsic ordering of these categories. These variables are used to name, categorize, or label the attributes. The region is an independent variable since its variations do not depend on another variable but the researcher's experimenting. It is a cause of change. The salary is a quantitative discrete variable because it is countable in a finite amount of time and has a countable number of values between any two values. It is a dependent variable because of its variations depending on the dependent variable. It is an effect of the change[1].

Analysis

For interpretations and calculation of significance level, statistical and practical significance, several Python packages were used, including pandas, numpy, matplotlib.pyplot and stats. Before the main calculation, the data was read using pandas. Table 1 shows the summary

[1] **#variables:** I identified and classified the types of relevant variables of the dataset and examined the relationship between them. I identified dependent and independent variables and classified them as quantitative and qualitative variables. It was a necessary step since it affects the type of statistical analysis and visualization that are appropriate.

statistics for Mid-career 90th percentile salary on dollars in Northeastern and South universities. The sample distributions of salaries for each region are displayed in Figure 1 and Figure 2. The full calculation is included in Appendix A.

	Mid-career 90th percentile salary, $	
	Northeast	Southern
Count	82	71
Mean	181926.83	152769.01
Median	173500.0	150000.0
Standard deviation	42439.14	32587.98
Range	209000	156700

Table 1. Summary statistics for the Mid-career 90th percentile salary in the Northeast and South

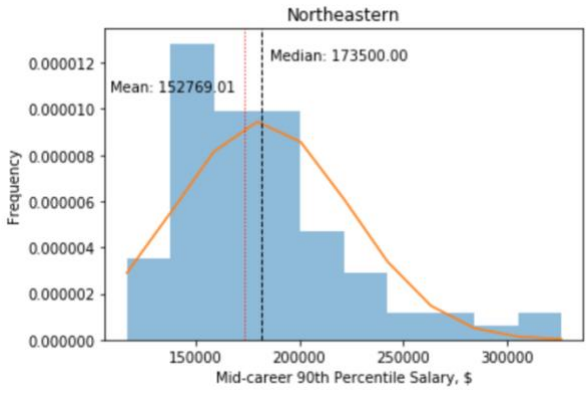


Figure 1. The sample distribution of salaries in the Northeast.

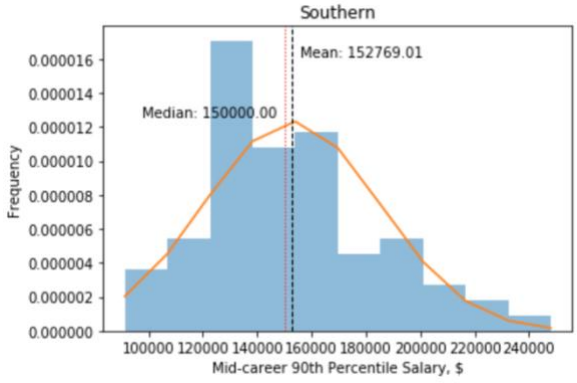


Figure 2. The sample distribution of salaries in the South

Figures 1 and 2 depict the sample distribution of the data. The Mid-career 90th Percentile Salary distribution in the Northeast is slightly positively skewed, with mean and median differing for $20000. There are some outliers on the right tail, which increase the median salary and lead to a higher standard deviation. In contrast, the southern salaries are almost normally distributed since the mean and median are very close, and there are not extreme values on tails. Figure 1 shows that there might be a wage gap between different occupations as the range of values is almost two times the minimum salary. In contrast, the southern salaries are less distributed with a smaller standard deviation[2].

In the analysis, we must use the t-distribution because we are working with the problem when the population standard deviation (σ of salaries from all universities in Northeastern and Southern US) is unknown, and the sample size is small relative to the whole population. There are approximately 1500 colleges in the Northeast region. Thus, the sample of 82 universities included in the data comprise less than 10% of the population. Similarly, 71 universities in the South also account for less than 10% of all existing institutions in the region (Study.com, 2020).

The t-distribution is used for inference when working with a difference of two means if each sample meets the conditions for using the t-distribution, and the samples are independent. To meet the first requirement, several conditions of t-distribution should be evaluated.

1. *Randomness*. We assume that the data come from a simple random sample. A random sample implies that we use a non-deterministic method to select a sample from the

[2] **#descriptivestats:** I calculated and interpreted descriptive statistics, including the mean, median, standard deviation, and range. It provided an overview of the key properties of the entire set of data. The calculation and interpretation of mean and median provided a better understanding of the skewness and potential significant outliers in the dataset. As a result, I derived Table 1 and Figures 1 and 2.

population. Here, we are looking for a sample where each population element has an equal probability of inclusion in a sample.

2. *Independence of observations.* Because the data come from a simple ransom sample and consist of less than 10% of the whole population, the observations are independent. Additionally, while the Northeastern sample is slightly right-skewed, the sample size of 82 would make it reasonable to model each means separately using a t-distribution. Ultimately, the data meets this criterion since we assume the independence if the sample is taken from less than 10% of all universities in the regions.
3. *Observation come from a nearly normal distribution.* According to histograms, the data is mainly might seem slightly skewed. Nevertheless, since both sample sizes are larger than 30, the t-distribution becomes more like a normal distribution. Thus, given the large small size, the distribution is allowed to be skewed.

Regarding the second condition of independence for using t-distribution for a difference of mean, the independence reasoning applied in the second condition above ensures that each sample's observations are independent. Since both conditions are satisfied, the confidence interval and hypothesis testing of the difference in sample means may be modeled using a t-distribution[3].

[3] **#distribution:** In this analysis, I gave reasons why the t-distribution is more appropriate than the z-score and normal distribution. Subsequently, I made inferences based on sample distribution to find confidence level, significance, and practical tests and test the hypothesis. I distinguished between population, sample, and sampling distributions and explained why I need to calculate the mean difference. I used the concept of the Central Limit Theorem to derive the sampling distribution and do my calculations. To do so, I also checked the requiring conditions for t-distribution for a difference of means and analyzed assumptions and consequences of the test's limitations.

First, we need to calculate the confidence interval for the difference of mid-career 90th percentile salaries for Northeastern and Southern, which provides a plausible range of values for possible differences. The goal is to identify a 95% confidence interval. Since we calculate the standard error from the sample standard deviation, we use the t-distribution for inference in our sampling distributions. Above we proved that all conditions are satisfied. Since these conditions are met, the Central Limit Theorem ensures that we can calculate our confidence interval from the normal distribution. A point estimate of the difference in the salary can be found using the difference in the sample means:

$$point\ estimate = \mu_{northeastern} - \mu_{southern}$$

We can quantify the variability in the point estimate using the following formula for its standard error. It is estimated using standard deviation estimates based on the sample:

$$SE_{\mu_{northeastern}-\mu_{southern}} = \sqrt{\frac{{\sigma_{northeastern}}^2}{n_{northeastern}} + \frac{{\sigma_{southern}}^2}{n_{southern}}}$$

Because we are using the t-distribution, we also must identify the degrees of freedom. A technique is to use the smaller of $n_{northeastern} - 1$ and $n_{southern} - 1$.

For a 95% confidence interval, we will use the sample difference and the standard error for that point of estimate from the earlier calculations. Using Python to find the t-score for 95% confidence interval and the derived degrees of freedom, we derive the following expression:

$$point\ estimate \pm t \times SE$$

The complete calculations of the confidence interval can be found in Appendix C[4]. Thus, we are 95% confident that the true population difference between mid-career 90th percentile salaries in the Northeast and South lies between \$17038.95 and \$41276.68. Since both lower and upper bounds are positive, it gives plausible evidence that Northeastern's 90th percentile salaries are larger than those in the South. In the context of university graduates, it means that the graduates from the Northeast will have larger salaries a couple of years later their graduation[5].

The confidence interval gives us a plausible range of difference of means of the sample. Nevertheless, to test the null hypothesis and address the research question, one must perform a test for statistical significance. We set the significance level (Type 1 Error) to $\alpha = 0.05$. It is the probability of rejecting the null hypothesis when it is true. In other words, there is a 5% risk of concluding that a difference exists when there is no actual difference. The previously introduced null hypothesis states that the two means are equal, while the alternative states that northeastern salaries are larger than southern. Thus, the test is one-tailed since we are only interested in the extreme values in the right tail (larger values). Since we are calculating the standard error using the sample standard deviation, we use the t-distribution. As we described before, all the conditions and assumptions for t-distribution are met and well justified.

---

[4] **#algorithms:** Throughout the work, I used the HC to implement working code and ensure that I am applying algorithmic thinking strategies to solve a problem and answer the research question. I had a set of steps that lead me to the desired output, given an input. The steps of the algorithm are well-ordered, clear, unambiguous, effective, and doable. I mostly worked with functions so that the algorithm is robust and can handle a range of inputs rather than a small set of specific cases. It is relatively simple and efficient with no unnecessary redundancies. It terminates in a finite number of steps. Finally, I added comments when appropriate to guide readers through the code.

[5] **#confidenceinterval:** I applied and interpreted confidence intervals with specified upper and lower values that are likely to include an unknown true population mean. The intervals have a 95% confidence level, which refers to the reliability of constructing the interval in the frequentist perspective. Thus, if the sampling procedure were repeated a lot of times, with distinct point estimates and confidence intervals obtained, we expect 95% of these intervals to contain the true population mean.

To assess statistical significance by computing the p-value, firstly, one should calculate T-score. The standard error is estimated in the previous calculation of the confidence intervals.

$$T = \frac{\mu_{southern} - \mu_{northeastern}}{SE}$$

Next, we compare the derived value in the t-table, where we use the smaller of $n_{northeastern} - 1$ and $n_{southern} - 1$ as the degrees of freedom. Using the stats package in Python, the estimate the p-value of 4.37*10^-6 (See Appendix D for the full calculations). The p-value is smaller than the significance level which is enough to reject the null hypothesis and accept the alternative one. Thus, there is statistically significant evidence that the mean of starting median salary for Northeastern schools is larger than that of Southern schools.

Next, to assess the practical significance, we need a measure of effect size. Statistical significance does not imply practical significance (or reversely) since p-values strongly depend on sample size.

$$d = \frac{\mu_{northeastern} - \mu_{southern}}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{\frac{(n_{northeastern} - 1) \times {\sigma_{northeastern}}^2 + (n_{southern} - 1) \times {\sigma_{southern}}^2}{n_{northeastern} + n_{southern} - 2}}$$

Both Cohen's d and Hedge's g pool variances on the assumption of equal population variations, but Hedge's g pools using n-1 for each sample instead of n, providing a better estimate. Although both effect sizes are positively biased, it almost negligible for the given

moderate sample sizes. We can reduce the bias of Cohen's d using the following formula for Hedge's g:

$$g = d \times (1 - \frac{3}{4(n_{northeastern} + n_{southern}) - 9}$$

Thus, we derive d=0.7635 and g=0.7597 (See Appendix E for detailed code and calculations in Python). These values are almost identical. It is reasonable since the sample sizes are larger than 20. We can get a better correction if the Hedge's g was calculated for smaller sample sizes[6].

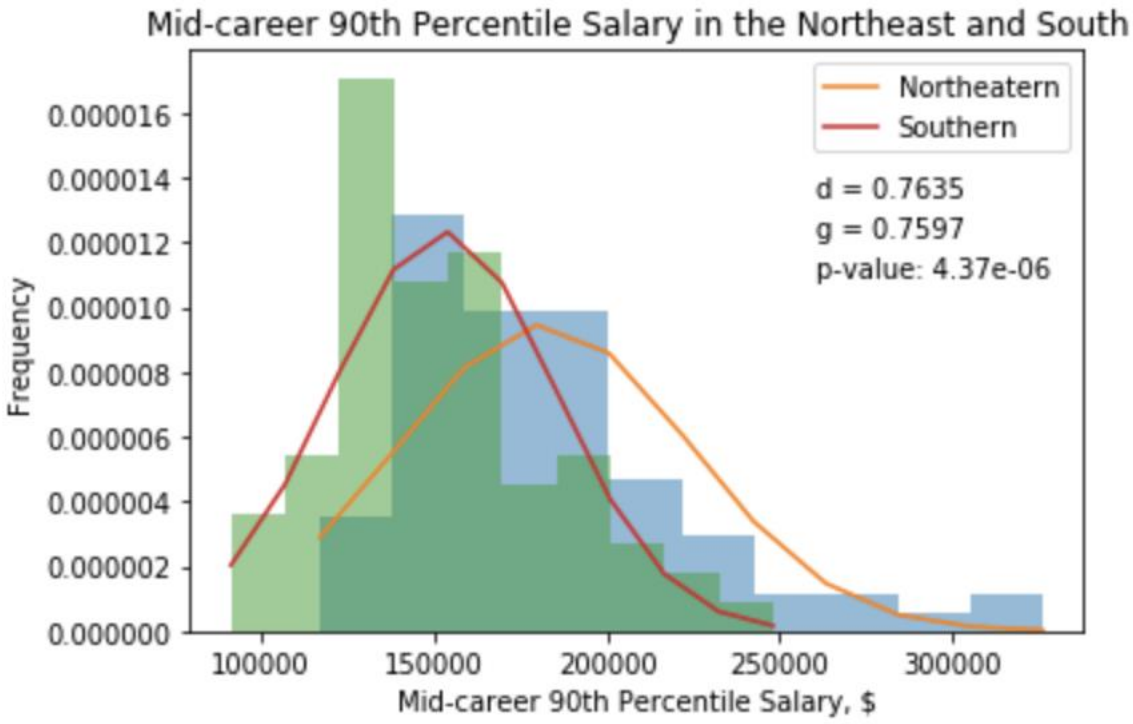


Figure 3. The final results

[6] **#significance:** I applied, interpreted, and distinguished practical and statistical significance. I computed a statistical significance test, which tells us when a difference in a sample is likely due to chance or reflects strong evidence of an actual difference. To know whether the derived difference matter, I did a test for practical significance, which gave me an effect size.

Thus, from the effect size, we can say that Northeastern salaries are significantly larger than Sothern's (Figure 3). In the context of analysis, it means that the true population means lie within 0.76 standard deviation from each other, which is a significant difference[7].

Results and Conclusion

We answered the initial research question by checking and justifying assumptions and applying different tests. From the 95% confidence interval of difference of means of Mid-career 90th Percentile Salaries in Northeastern and Southern states (17038.95, 41276.68), it is evident that the difference is positive and that generally Northeastern salaries are higher. Further, estimating the statistical significance showed that the p-value is 4.37*10^-6. It implies that the probability of obtaining an effect at least as extreme as the one in the sample data is very low, assuming the truth of the null hypothesis. Thus, the samples provide enough evidence that the null hypothesis for the entire population can be rejected. Next, the practical significance derived using the Hedge's g demonstrates that the means of two populations are 0.76 standard deviation from each other which is a large effect. Thus, we accept the alternative hypothesis and conclude that there is statistically significant evidence that the mean of starting median salary for Northeastern schools is larger than that of Southern schools.

We came to this conclusion using inductive reasoning, because we form likely generalizations about the population based on specific incidents and calculations. The premises are the tests that were conducted. They gave more plausible conclusions about the

[7] **#dataviz:** I effectively generated a detailed data visualization appropriate for the final calculated values. I analyzed and interpreted the patterns (Figure 3).

difference of mean. Nevertheless, the results are not fully reliability since we worked only with a small fraction of all population values[8].

Overall, there is statistically significant evidence that the mean of starting median salary for Northeastern schools is larger than that of Southern schools. To further investigate this matter, one can collect data and research using questionaries to assess other variables that might influence the salaries, including the people's age, years of experience, dropout rates of universities, and the graduate degree. Subsequently, one can apply a t-test to find the difference of means between these variables. The following research questions can be asked: Is there convincing evidence that graduates with two years of experience earn more than employees with three years of experience?" or "Is there convincing evidence that people with a graduate degree in mathematics earn more than graduate students with a physics degree?"[9]

[8] **#induction:** I analyzed and applied inductive reasoning using evidence to support my conclusion and reasoning from a set of instances to a generalization about them. The conclusion of the analysis goes beyond the content of premises since I create a general claim about the population. Thus, the truth value of the argument cannot be guaranteed. I evaluated the results' strength and reliability, assessing how likely the conclusion is given that the premises are true.

[9] **#professionalism:** I followed established guidelines for presenting my work professionally. To communicate effectively, I considered my tone and forms of address. I proofread the final written work for errors several times and adequately attributed quotations, ideas, and other sources. I used APA citations for citations and formatting of my bibliography. I have several scholarly and popular sources. Overall, I ensured that my approach to communication meets the expectations relevant to the context.

References[10]

Biddix, J. P. (2009). *Effect Size.* Research Rundowns. Retrieved January 26, 2016, from https://researchrundowns.wordpress.com/quantitative-methods/effect-size/

Diez, D., Barr, C., & Cetinkaya-Rundel, M. (2015). Sections 4.3 and 6.1 in *OpenIntro Statistics (3rd ed.)*.

Lane, D. (n.d.). *Logic of Hypothesis Testing*. OnlineStatBook.

Study.com. (2020). List of Colleges in the Northeast U.S. Retrieved 13 December 2020, from https://study.com/colleges_in_the_northeast.html

The Wall Street Journal. (2020). Where it Pays to Attend College. Retrieved 12 December 2020, from https://www.kaggle.com/wsj/college-salaries

Reflection

In this paper, I believe I improved my #induction skills. I evaluated the strength and reliability of the argument. One of my assignment feedback stated that I mistakenly named the Wilkins factors. Here, we use an eliminative induction in which you present a variety of different pieces of evidence that all support the conclusions. In this context, I developed and applied several tests to challenge my null hypothesis. In y particular example, this induction method means that I gradually gain more confidence that there is enough evidence to reject the null and accept the alternative one[11].

[10] **#sourcequality:** To determine the source quality, I distinguished between categories and types of information. I ordered my sources based on their relevance, currency, accuracy, authority, and purpose. Subsequently, I chose the ones that best fit the purpose of my work based on this categorization.

[11] **#organization:** To effectively communicates with readers, I organized my work following the Technical Report format. Firstly, I had an introduction that explained what this analysis is about. Then, I gave some information about the dataset, variables, and statistics summary. Subsequently, I found a confidence interval for my difference of means and calculated several tests. Eventually, I concluded the final estimations of inference and reflected on the results.

Appendix A: Summary Statistics

```
#the modification of the code from Session 24 CS50

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import matplotlib
from scipy import stats

# defining a function to print out some simple statistics of the data
#it is useful for #varibales and the further caluclations of tests for␣
 ↪statistical and practical significance
def print_stats(list):
    print('count:',len(list))
    print('mean:',np.mean(list))
    print("median:", np.median(list))
    print("range:", max(list)-min(list))
    print('std:',np.std(list,ddof=1),"\n") #Bessel's correction

salariesdata = pd.read_csv('salaries-by-region.csv')

#extracting the right column (7th column) for mid-career 90th percentile salary␣
 ↪based on the region
northeastern = list(salariesdata[salariesdata.Region == 'Northeastern'].values[:
 ↪,7])
southern = list(salariesdata[salariesdata.Region == 'Southern'].values[:,7])

Southern = []
Northeastern = []

#the salary entries have $ signs, commas and dots that makes it harder to␣
 ↪process the data
#thus, it was decided to erase them
for i in southern:
    new = i.replace("$", "")
    n = new.replace(".00", "")
    m = n.replace(",","")
    Southern.append(int(m))
```

```
for i in northeastern:
    new = i.replace("$", "")
    n = new.replace(".00", "")
    m = n.replace(",","")
    Northeastern.append(int(m))

# print the stats for each category
print('Northeastern')
print_stats(Northeastern)

print('Southern')
print_stats(Southern)
```

Results:

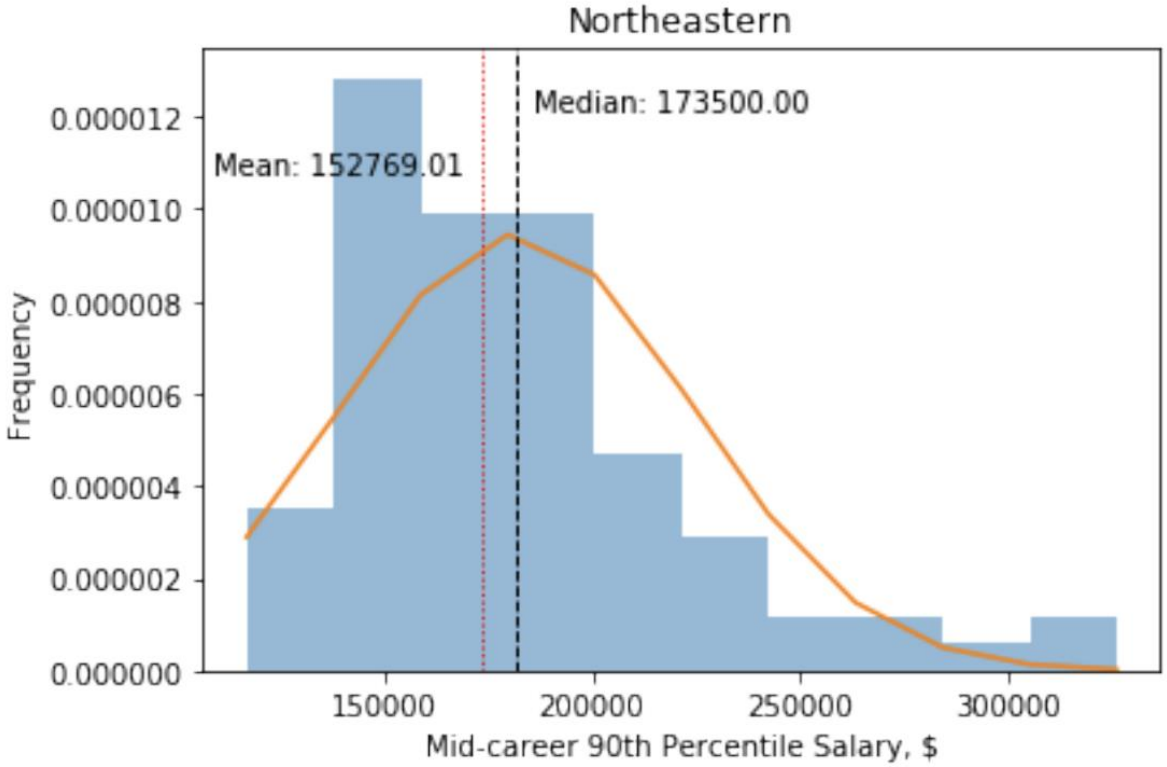
```
Northeastern
count: 82
mean: 181926.82926829267
median: 173500.0
range: 209000
std: 42439.14466266828

Southern
count: 71
mean: 152769.01408450704
median: 150000.0
range: 156700
std: 32587.980218607754
```

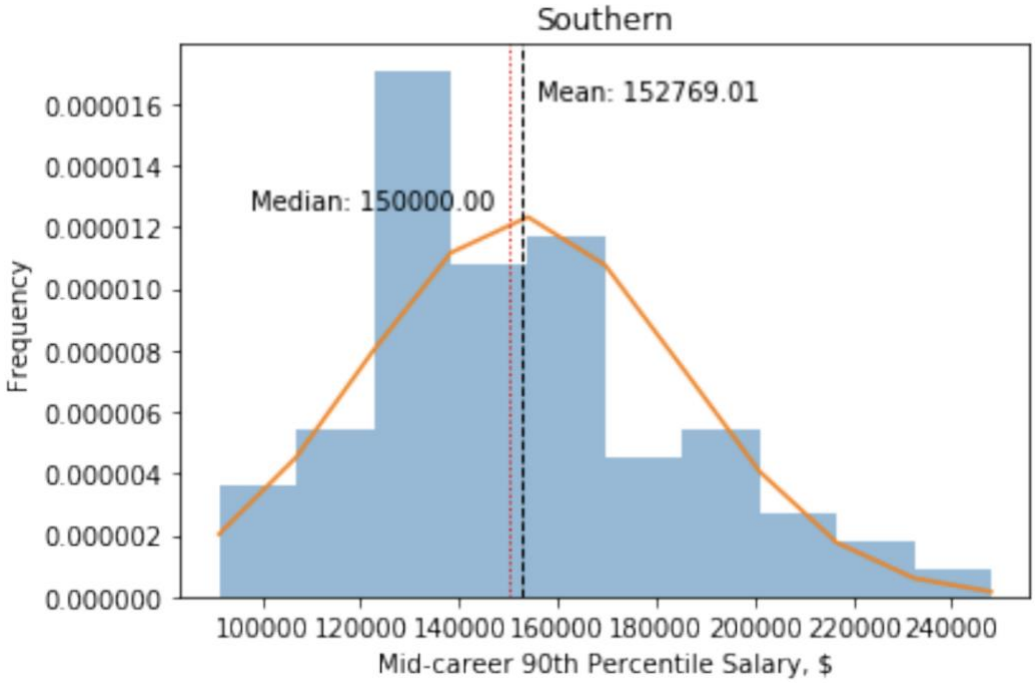
Appendix B: #descriptivestats, #dataviz

```
#drawing historgrams for #descriptivestats and assessing conditions for the test

#plt.hist(Northeastern)
_, bins, _ = plt.hist(Northeastern, 10, density=1, alpha=0.5)
mu, sigma = stats.norm.fit(Northeastern)
best_fit_line = stats.norm.pdf(bins, mu, sigma)
plt.plot(bins, best_fit_line) #drawing normal curve
plt.title('Northeastern')
plt.xlabel('Mid-career 90th Percentile Salary, $')
plt.ylabel('Frequency')
min_ylim, max_ylim = plt.ylim()
plt.axvline(np.mean(Northeastern), color='k', linestyle='dashed', linewidth=1)
plt.text(np.mean(Northeastern)*0.6, max_ylim*0.8, 'Mean: {:.2f}'.format(np.
 ↪mean(Southern)))
plt.axvline(np.median(Northeastern), color='r', linestyle='dotted', linewidth=1␣
 ↪)
plt.text(np.median(Northeastern)*1.07, max_ylim*0.9, 'Median: {:.2f}'.format(np.
 ↪median(Northeastern)))
```

```
plt.show()

#plt.hist(Southern)
_, bins, _ = plt.hist(Southern, 10, density=1, alpha=0.5)
mu, sigma = stats.norm.fit(Southern)
best_fit_line = stats.norm.pdf(bins, mu, sigma)
plt.plot(bins, best_fit_line) #drawing normal curve
plt.title('Southern')
plt.xlabel('Mid-career 90th Percentile Salary, $')
plt.ylabel('Frequency')
min_ylim, max_ylim = plt.ylim()      #deriving the upper and lower bounds of␣
 ↪the y axis, which is necessery for putting a text next to the vertical lines
plt.axvline(np.mean(Southern), color='k', linestyle='dashed', linewidth=1)   ␣
 ↪#drawing a vertical line representing previously derived mean
plt.text(np.mean(Southern)*1.02, max_ylim*0.9, 'Mean: {:.2f}'.format(np.
 ↪mean(Southern)))    #putting a text with the corresponding value of the mean
plt.axvline(np.median(Southern), color='r', linestyle='dotted', linewidth=1 ) ␣
 ↪#another vertical line showing the median
plt.text(np.median(Southern)*0.65, max_ylim*0.7, 'Median: {:.2f}'.format(np.
 ↪median(Southern)))   #labeling the median
plt.show()
```

Results:


Northeastern
Median: 173500.00
Mean: 152769.01
Frequency
0.000012
0.000010
0.000008
0.000006
0.000004
0.000002
0.000000
150000
200000
250000
300000
Mid-career 90th Percentile Salary, $


Southern
Mean: 152769.01
Median: 150000.00
Frequency
0.000016
0.000014
0.000012
0.000010
0.000008
0.000006
0.000004
0.000002
0.000000
100000 120000 140000 160000 180000 200000 220000 240000
Mid-career 90th Percentile Salary, $

Appendix C

```
from scipy import stats

#fisrt, we calcualate the confidence interval
mean = np.mean(Northeastern) - np.mean(Southern) #difference of mean
SE = (np.std(Southern,ddof=1)**2/len(Southern)+np.std(Northeastern,ddof=1)**2/
 ↪len(Northeastern))**0.5   #standard error of difference of mean with the␣
 ↪Bessel's correction applied
degrees_of_freedom = min(len(Northeastern)-1, len(Southern)-1) #the degree of␣
 ↪freedom is the minimum of two sample sizes minus one: conservative estimate␣
 ↪from OpenIntro
t = stats.t.ppf(0.975,degrees_of_freedom) #calculating the t-score from p; 0.
 ↪975 bcause we added 2.5% on the left tail
print("Confidence interval: [", mean - t * SE, ",", mean + t* SE, "]")␣
 ↪#calculation of the lower and upper bound of the interval
```

Results:

```
Confidence interval: [ 17038.95203995214 , 41276.67832761911 ]
```

Appendix D

```
#the calcualtion of the p-value
T = (np.mean(Southern)-np.mean(Northeastern))/SE #finding the t-score
print("p-value:", stats.t.cdf(T,degrees_of_freedom))
```

Results:

```
p-value: 4.369697827622665e-06
```

Appendix E

```
: #assessing the practical significance
SDpooled = np.sqrt((np.std(Northeastern,ddof=1)**2*(len(Northeastern)-1) + np.
 ↪std(Southern,ddof=1)**2*(len(Southern)-1))/
 ↪(len(Northeastern)+len(Southern)-2)) # OpenIntro section 5.3.6
Cohensd = mean/SDpooled
Hedgesg = Cohensd * (1-3/(4*(len(Northeastern)+len(Southern))-9))
print("Cohen's d:", Cohensd)
print("Hedge's g:", Hedgesg)
```

Results:

```
Cohen's d: 0.7635003546675959
Hedge's g: 0.759701845440394
```

Appendix F

```
_, bins, _ = plt.hist(Northeastern, 10, density=1, alpha=0.5)
mu, sigma = stats.norm.fit(Northeastern)
best_fit_line = stats.norm.pdf(bins, mu, sigma)
plt.plot(bins, best_fit_line, label="Northeatern") #drawing normal curve
plt.title('Mid-career 90th Percentile Salary in the Northeast and South')
plt.xlabel('Mid-career 90th Percentile Salary, $')
plt.ylabel('Frequency')

_, bins, _ = plt.hist(Southern, 10, density=1, alpha=0.5)
mu, sigma = stats.norm.fit(Southern)
best_fit_line = stats.norm.pdf(bins, mu, sigma)
plt.plot(bins, best_fit_line, label="Southern") #drawing normal curve

min_ylim, max_ylim = plt.ylim()      #deriving the upper and lower bounds of␣
 ↪the y axis, which is necessery for putting a text next to the vertical lines
min_xlim, max_xlim = plt.xlim()

plt.text(max_xlim*0.77, max_ylim*0.75, 'd = 0.7635')    #putting a text with␣
 ↪the resulting statistics
plt.text(max_xlim*0.77, max_ylim*0.68, 'g = 0.7597')
plt.text(max_xlim*0.77, max_ylim*0.61, 'p-value: 4.37e-06 ')

plt.legend()
plt.show()
```

Results: