Inferential Statistics Report

Introduction

Statistics about salaries are an interesting topic, especially those paid by colleges. In this report, we explore how the median starting salaries paid by colleges vary among different regions (Midwestern and Northeastern). We first process the data by calculating the descriptive statistics and computing the distribution of two sample groups. Then, we construct a confidence interval to estimate the average median starting salaries in different colleges. We also perform a test for statistical and practical significance to determine whether professors in the Midwestern region are better-paid than those in the Northeastern region. In other words, is there convincing evidence that colleges pay teachers differently in two areas of the United States? We really want to see if there is enough convincing evidence to conclude that colleges in one region pay better than those in the other.

Dataset

The dataset contains several kinds of salaries categorized by several different ways. However, as I am interested in two specific regions, Midwestern and Northeastern, as well as the median starting salaries, I decided to filter data to make a comparison between these two samples. The school names are not important for this analysis, so one variable of interest is the median starting salary, a quantitative variable. To specify, salary can actually be both continuous and discrete, depending on the context. In this dataset, from my point of view, the variable is discrete because although it involves decimals in salaries, in reality, no one is paid $100,520.3134 for example. Therefore, we can still pick a specific number for a salary, which goes against a continuous variable. Another variable involved in this analysis is region. This is a categorical qualitative variable which serves to categorize salaries in colleges of two different areas.

Analysis

Summary Statistics

The dataset was processed in Python. We first calculated the descriptive statistics for two quantitative variables, salaries paid by Northeastern and Midwestern colleges, including mean, median, mode, range, and standard deviation. Table 1 provides the summary statistics for median starting salaries in both Midwestern and Northeastern colleges (these are computed in Appendix A). The sample distributions for each quantity are displayed in Figures 1 and 2 (these are computed in Appendix B).

Table 1: Summary statistics for the variables of interest		
	Starting Median Salaries in Midwestern Colleges	Starting Median Salaries in Northeastern Colleges
Count	$n = 64$	$n = 82$
Mean	$\bar{x}_m = 44,461$	$\bar{x}_n = 48,679$
Median	43,200	46,700
Mode	39,800	45,700
Standard Deviation	$s_{age} = 5,149$	$s_{age} = 7541$
Range	21,300	35,300

Figure 1. Histogram for median starting salaries in Midwestern colleges

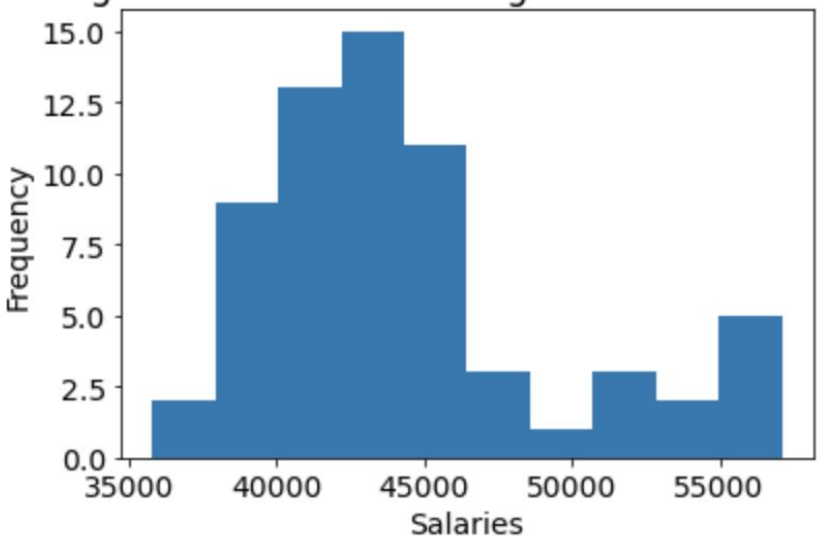
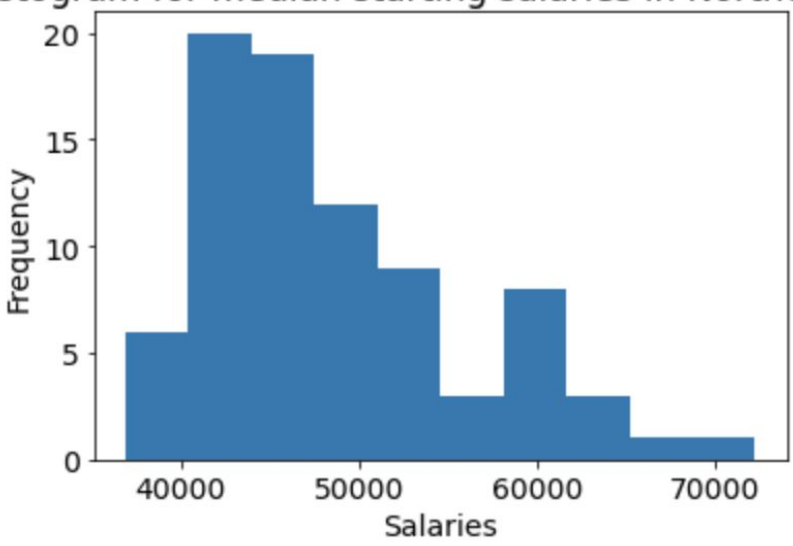


Figure 2. Histogram for median starting salaries in Northeastern colleges



From the two histograms, we can understand that both sampling distributions are right-skewed, not normal, because there are more values in the left end of the distribution. However, from the descriptive statistics, we can evaluate this skewness and justify that there are no significant outliers. To specify, the ranges of two sample groups are 21,300 and 35,300, respectively. Meanwhile, the means are more than 44,000 and 48,000, correspondingly, which does not show any big differences. Moreover, the medians are quite close to the means (43,200 and 46,700). For that reason, we can learn that two distributions are not largely right-skewed, which is important in the distribution analysis below because if combined with meeting conditions, these two distributions are eligible for applying the Central Limit Theorem to generate approximately normal distributions.[1][2]

Confidence Interval

We then compute a confidence interval for the median starting salaries paid by Midwestern colleges. We choose a confidence level of 95%. This 95% confidence interval will provide us with a plausible range of values for starting salaries. Since we do not have data of salaries paid by all colleges in two regions, we do not know the population standard deviation. Moreover, we are using on means. Therefore, we choose to use the t-distribution for inference and estimate the standard error using the sample standard deviation. Furthermore, we also use Bessle's correction to the standard deviation to

[1] **#descriptivestats**: I chose an appropriate set of descriptive statistics serving to evaluate the distribution of two sample groups. I also calculated them with great accuracy and clearly stated steps in the algorithms. The interpretation I provided also worked its efficiency to justify the distributions' eligibility for applying the Central Limit Theorem.
[2] **#dataviz**: Generated a suitable data visualization (histogram to understand the distributions of sample groups). Interpreted this data visualization detailedly to prepare for below tests and calculations.

correct the bias in the estimation of the population variance. To ensure this choice works well, we must check if three conditions are met:

- Random: The colleges were randomly selected to investigate on their salaries paid
- Normal: n > 30. This condition holds because both sample groups have sample size larger than 30 (64 and 84 specifically)
- Independent: We check this condition because n < 10% of population size.

Since these conditions are met, we are justified to use the t-distribution. As a result, we can totally apply the Central Limit Theorem for all samples to be normally distributed when the number of samples gets larger.[3]

Coming to the calculation part, we choose to construct a 95% confidence interval for the median starting salaries paid by Midwestern colleges. We choose t-score instead of z-score because population standard deviation is unknown and we need greater exactitude. We need to use degrees of freedom instead of sample size to make our results more accurate, using the formula: $df = n - 1$. With confidence level and degrees of freedom, we can find out the t-score thanks to the calculator. We then compute the standard error using the usual formula: $SE = \frac{SD}{\sqrt{df}}$. The sample standard deviation is known, thus facilitating us to do the calculation. The margin of error would be the product of t-score and the standard error. The full calculation can be found in Appendix C.[4] The resulting 95% confidence interval for the mean ice-cream consumption is (43,164.5, 45757.4), which provides a plausible range of values for the median starting salaries paid by Midwestern colleges.

For interpretation of this confidence interval, *we can be 95% confident that the population mean of median starting salaries paid by Midwestern colleges will fall into the interval (43,164.5, 45757.4),* which means that *if we conduct taking sample with the same sample size several times, 95% of the confidence intervals we construct from these samples will contain the population mean.* In this context, we can understand the range of median starting salaries in Midwestern colleges.

Difference of Means Test

We also want to make a comparison between two sample groups, Midwestern and Northeastern. To address the question of median starting salaries vary in two different regions, we perform a difference of means significance test. We choose a t-test because we want to compare the means of two groups

[3] **#distributions**: This is a strong application of this HC because I identified and justified a kind of distribution of my sampling distributions by checking all conditions carefully. Additionally, I described accurate outstanding features of this distribution detailedly.

[4] **#confidenceintervals**: This is a strong application of this HC because I calculated several components and successfully constructed an appropriate confidence interval with clearly stated steps. Also, I interpreted exactly the meaning of the confidence interval and applied this into following tests to examine the population.

with great exactitude. We set our significance level to the default, $\alpha = 0.05$. We also clearly define both the null and alternative hypotheses:

- Null hypothesis: $x_M = x_N$
- Alternative hypothesis: $x_M \neq x_N$

"M": the median starting salaries paid by colleges in Midwestern regions
"N": the median starting salaries paid by colleges in Northeastern regions

Note that the test is 2-tailed because we are looking for a difference in either direction.

Although both sample sizes are large enough ($n_1, n_2 > 30$), because we are estimating the standard error using the sample standard deviations, we use the t-test instead of z-test for the significance test. As shown in Table 2, these are specific statistics of two sample groups we mainly use in the test.

Table 2: Summary statistics for the starting median salaries (in dollars) for two sample groups: Midwestern colleges and Northeastern colleges

	Midwestern	Northeastern
Count	$n_1 = 64$	$n_2 = 82$
Mean	$\bar{x}_1 = 44460.9$	$\bar{x}_2 = 48679.3$
Standard Deviation	$s_1 = 5149.4$	$s_2 = 7541.3$

Regarding statistical significance, p-value is our ultimate goal. To de so, we first compute the T-score using the usual formula for a difference of means test, $T = \frac{\bar{x}_2 - \bar{x}_1}{SE}$ with $SE = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$.

We choose the smaller degrees of freedom to best do the calculation, which is $df = n_1 - 1$. See Appendix D for the calculation in Python.

The T-score of 4.01 results in a two-tailed p-value of 0.0002 < 0.05. The probability is too small, meaning that it is very unlikely that we can get such a sample for the null hypothesis to be true. Therefore, we conclude that the data favours rejecting that the mean of median starting salaries is similar in two regions.

Regarding practical significance, we need to find out the effect size. In this case, we choose Hedge's g as our measure because we want to acquire more accurate results and eliminate the upward

bias that Cohen's d has. Computing Hedge's requires the pooled standard deviation which can be figured out by the formula: $SDpooled = \sqrt{\frac{(s_1^2 + s_2^2)}{2}}$. The calculation can be shown in Appendix D, producing an effect size of 0.66. Therefore, we can say that the difference is not trivial, yet actually nearly large because two groups' means differ by 0.66 (> 0.50) standard deviations. In this particular case, we can say Northeastern and Midwestern colleges offer quite different median starting salaries.[5]

Results and Conclusions

Throughout the report of examining the data about median starting salaries paid by Midwestern and Northeastern colleges, we successfully constructed a 95% confidence interval for the mean median starting salaries in Midwestern colleges, (43,164.5, 45757.4) (outputted in Appendix C). From that, we can learn that with the same sample size, if we take out several samples and construct one confidence interval correspondingly, 95% of the confidence intervals will contain the population mean. Furthermore, the test of statistical and practical significance also sheds insights into two null and alternative hypotheses. We computed the results, p=0.0002 and g=0.66, which results in convincing evidence that there is a difference between two groups' means and it is not only statistically significant but also quite large in practice. In this context, this suggests that mean median starting salaries in both Midwestern and Northeastern colleges differ significantly.

These conclusions are supposed to be inductive because we only examine the data about samples, and process the data to gain insights into statistics of two sample groups. Based on sample examination, we reached the conclusion about population groups. Therefore, the conclusion is not contained within the scope of premises and we are using the method of generalization. We still need more samples with larger sample sizes for the sake of greater accuracy. However, they are strong because they are built on appropriate assumptions which are conditions holding, and large enough sample size (n > 30). The p-value also supports the strength of this inductive reasoning because it is really small, which rejects the null hypothesis. The reliability of this induction is good with a quite large effect size, yet can still be improved if effect size is more than 0.80.[6]

In summary, we've shown that there is statistically significant evidence that the mean of median starting salary for Midwestern schools is larger than that of Northeastern ones.

Reflection

[5] **#significance**: This is a strong application of this HC because I applied an accurate and appropriate significance test with justified reasoning. I also distinguished well between practical and statistical significance to fully justify the results of the test.

[6] **#induction**: This is an awesome application of this HC because I accurately and effectively identified inductive reasoning and justified it with inductive characteristics. I also specified the type of induction and explained it carefully

I acknowledge the valuable feedback from the Professor about whether the variable of money should be determined to be continuous or discrete, his advice on focusing on the context, and his expectation on students to justify their opinions because there are no right answers. This truly motivated me to delve deeper into my research and engagement with the data so that I can understand the function of this variable in this particular case. Based on this, my application of #variables is hopefully satisfactory enough.

Appendix

The full Jupyter notebook file can be accessed here [http://localhost:8888/notebooks/Downloads/Test.ipynb#]. The data can be accessed here.

Appendix A: Import and Analyze Data

```
#import relevant packages and 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 using pandas
#this reads the data into a "dataframe"
data = pd.read_csv("midwestern - Sheet1.csv", sep=',')
data.head(100)

```

	School Name	Starting Median Salary
0	University of Notre Dame	$56,300.00
1	University Of Chicago	$53,400.00
2	Illinois Institute of Technology (IIT)	$56,000.00
3	Case Western Reserve University	$56,200.00
4	University of Illinois at Urbana-Champaign (UIUC)	$52,900.00
...	...	...
59	Kent State University	$38,700.00
60	University of Wisconsin (UW) - Green Bay	$35,800.00
61	Indiana Wesleyan University (IWU)	$39,800.00
62	Pittsburg State University	$40,400.00
63	Davenport University	$39,700.00

64 rows × 2 columns

```
midwestern_list =  data['Starting Median Salary'].tolist()
mode_midwestern = stats.mode(midwestern_list)[0][0]
mean_midwestern = data["Starting Median Salary"].mean()
median_midwestern = data["Starting Median Salary"].median() # your code here
sd_midwestern = data["Starting Median Salary"].std() # your code here

print("- Median =",median_midwestern)
print("- Standard Deviation =",sd_midwestern)
print("- Mode =",mode_midwestern)
print("- Mean =",mean_midwestern)
print("- Range =", max(midwestern_list) - min(midwestern_list))


```

```
- Median = 43200.0
- Standard Deviation = 5149.4372380919085
- Mode = 39800.0
- Mean = 44460.9375
- Range = 21300.0
```

```
df = pd.read_csv("northeastern - Sheet1.csv", sep=',')
df.head(100)

```

	School Name	Starting Median Salary
0	Dartmouth College	$58,000.00
1	Princeton University	$66,500.00
2	Massachusetts Institute of Technology (MIT)	$72,200.00
3	Yale University	$59,100.00
4	Harvard University	$63,400.00
...	...	...
77	State University of New York (SUNY) at Potsdam	$38,000.00
78	Niagara University	$36,900.00
79	State University of New York (SUNY) at Fredonia	$37,800.00
80	University of Southern Maine	$39,400.00
81	Mercy College	$43,700.00

82 rows × 2 columns

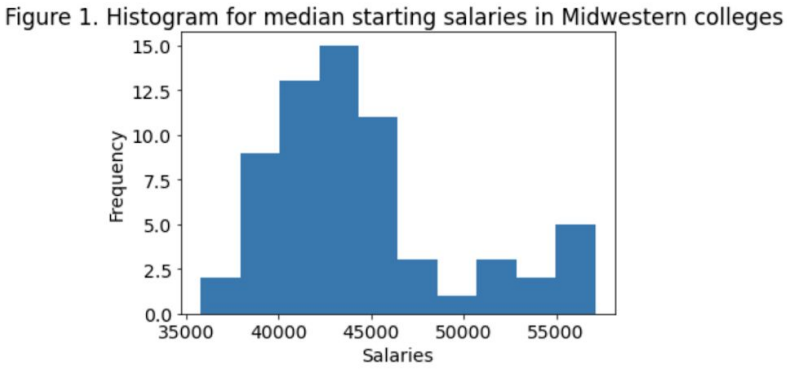
```
northeastern = list(df['Starting Median Salary'].values)
df["Starting Median Salary"] = df["Starting Median Salary"].str.replace("$","").str.replace(",","").astype(float
northeastern_list =  df['Starting Median Salary'].tolist()
mode_northeastern = stats.mode(northeastern_list)[0][0]
print("- Mode =",mode_northeastern)
mean_northeastern =  df['Starting Median Salary'].mean()
print("- Mean =",mean_northeastern)
median_northeastern = df['Starting Median Salary'].median() # your code here
sd_northeastern = df['Starting Median Salary'].std() # your code here

print("- Median =",median_northeastern)
print("- Standard Deviation =",sd_northeastern)
print("- Range =", max(northeastern_list) - min(northeastern_list))


```

```
- Mode = 45700.0
- Mean = 48679.26829268293
- Median = 46700.0
- Standard Deviation = 7541.322178420395
- Range = 35300.0
```

Appendix B: Visualize Data

```
midwestern = list(data['Starting Median Salary'].values)


print(midwestern)


#create histograms for each variable

plt.hist(midwestern, bins = 10)
plt.title("Figure 1. Histogram for median starting salaries in Midwestern colleges ")
plt.xlabel("Salaries")
plt.ylabel("Frequency")
plt.show()

```

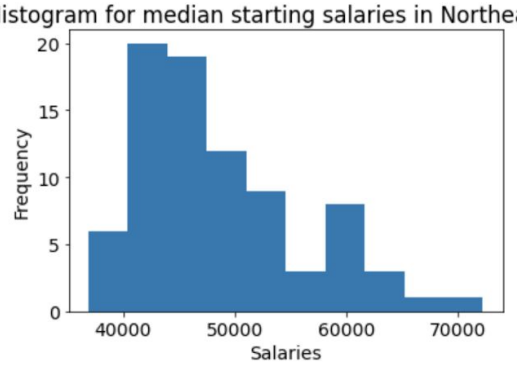
```
[56300.0, 53400.0, 56000.0, 56200.0, 52900.0, 52700.0, 57100.0, 55800.0, 52700.0, 51400.0, 48500.0, 41400.0, 48900.
0, 47000.0, 45300.0, 46300.0, 46400.0, 45400.0, 46200.0, 46300.0, 44700.0, 44900.0, 47500.0, 42400.0, 41700.0, 4570
0.0, 43600.0, 44000.0, 43300.0, 45800.0, 45100.0, 42800.0, 43100.0, 40800.0, 43300.0, 42300.0, 42300.0, 41100.0, 42
200.0, 42000.0, 43500.0, 41500.0, 43000.0, 40300.0, 39800.0, 41800.0, 40700.0, 41100.0, 36100.0, 42200.0, 43600.0,
38500.0, 41400.0, 39300.0, 38900.0, 41400.0, 39100.0, 44300.0, 39800.0, 38700.0, 35800.0, 39800.0, 40400.0, 39700.0
]
```



```
northeastern = list(df['Starting Median Salary'].values)
print(northeastern)
plt.hist(northeastern, bins = 10)
plt.title("Figure 2. Histogram for median starting salaries in Northeastern colleges ")
plt.xlabel("Salaries")
plt.ylabel("Frequency")
plt.show()
```

```
[58000.0, 66500.0, 72200.0, 59100.0, 63400.0, 60900.0, 62400.0, 61000.0, 61800.0, 61100.0, 60300.0, 54100.0, 56200.
0, 52800.0, 59400.0, 53900.0, 60600.0, 59200.0, 52700.0, 55800.0, 47800.0, 46500.0, 53000.0, 53600.0, 47200.0, 5030
0.0, 49000.0, 53400.0, 48000.0, 47300.0, 49500.0, 46800.0, 44500.0, 50300.0, 43800.0, 53200.0, 48900.0, 43600.0, 48
000.0, 48000.0, 46600.0, 42700.0, 45400.0, 50500.0, 45700.0, 49900.0, 43900.0, 45500.0, 43400.0, 48900.0, 40700.0,
47300.0, 44000.0, 52900.0, 44400.0, 41100.0, 44800.0, 46200.0, 42300.0, 42800.0, 45100.0, 42100.0, 40600.0, 45700.0
, 41800.0, 45600.0, 38000.0, 43200.0, 37500.0, 43000.0, 40800.0, 46000.0, 45700.0, 42100.0, 42400.0, 42000.0, 41200
.0, 38000.0, 36900.0, 37800.0, 39400.0, 43700.0]
```

Figure 2. Histogram for median starting salaries in Northeastern colleges



Appendix C: Confidence Interval

```
import math
import scipy
from scipy import stats
def confidenceintervals(mean, SD, n):
    q = 0.975 #q = (1-alpha/2) and alpha = 1 - confidence level
    df = n - 1
    SE = SD/(df**0.5)
    t_score = scipy.stats.t.ppf(q, df)
    print(t_score)
    lower_bound = mean - t_score*SE
    upper_bound = mean + t_score*SE
    print("The 95% confidence interval for salaries paid by Midwestern colleges is", [lower_bound,upper_bound])

confidenceintervals(mean_midwestern, sd_midwestern, 64)
```

```
1.9983405417721956
The 95% confidence interval for salaries paid by Midwestern colleges is [43164.477882252846, 45757.397117747154]
```

Appendix D: Difference of Means Test

```
1
2  def difference_of_means_test(data1, data2, tails):
3      n1 = len(data1)
4      n2 = len(data2)
5
6      x1 = np.mean(data1)
7      x2 = np.mean(data2)
8
9      s1 = np.std(data1,ddof=1)#Bessel's correction: use n-1 in denominator
10     s2 = np.std(data2,ddof=1)
11
12     SE = np.sqrt(s1**2/n1 + s2**2/n2)
13     Tscore = np.abs((x2-x1)/SE)
14     dof = min(n1,n2) - 1
15     pvalue = tails*stats.t.cdf(-Tscore,dof)
16
17     SDpooled = np.sqrt((s1**2*(n1-1)+s2**2*(n2-1))/(n1+n2-2))
18     Cohensd = (x2-x1)/SDpooled
19     Hedgesg = Cohensd*(1-(3/(4*(n1-n2)-9)))
20
21     print('p=', pvalue)
22     print("p={:f}".format(pvalue))
23     print('g=', Hedgesg)
24     print('t-score =',Tscore)
25
26     alpha = 0.05
27     if pvalue > alpha:
28         print("We are unable to reject the null hypothesis due to lack of evidence")
29
30     else:
31         print("We have enough convincing evidence to reject the null hypothesis")
32
33
34 difference_of_means_test(midwestern,northeastern,2)
```

```
p= 0.00016472973405458045
p=0.000165
g= 0.6625749208831034
t-score = 4.007693445164222
We have enough convincing evidence to reject the null hypothesis
```