# STAT 207 Homework 8 [50 points] - Solutions

## Hypothesis Testing and Difference Parameters

Due: Friday, March 24 by noon (11:59 am) CST

---

**Note:** For this assignment, you should not use any built-in functions for calculating confidence intervals or performing hypothesis tests. You will not earn credit if you use the built-in confidence interval or hypothesis test functions on this assignment. You may use other functions that we have discussed in class so far, including the mean, standard deviation, and sample functions.

## Package Imports [1 point]

Add code to the cell below to import packages needed for this assignment.

You may also need to read in additional packages later.

In [1]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm, t
```

## Case Study 1: Colleges and Universities (continued)

In Homework 7, you completed Case Study 2 about colleges and universities. In this Homework, we will return to the same data to complete additional inference procedures.

We will use a random sample of post-secondary education facilities (colleges & universities) from the United States, including Puerto Rico & other US territories. The **colleges.csv** file contains a random sample of 135 post-secondary education facilities from the US. This data originally came from https://www.kaggle.com/yamqwe/colleges-and-universitiese, although the data is no longer accessible through the web.

We will use three variables in particular for this assignment:

- the number of students enrolled full-time ( `FT_ENROLL` )
- the number of students enrolled part-time ( `PT_ENROLL` )
- the number of people employeed by the school ( `TOT_EMPLOY` )

## 1. Read in the data [2 points]

Below, you should read in the data from the **colleges.csv** file, perform any cleaning that needs to take place, and create the following variable:

- the variable `majority_pt` , a logical (Boolean) variable that indicates if the college or university has more students enrolled part-time than full-time (using `FT_ENROLL` and `PT_ENROLL` )

**Hint:** For the data cleaning, you may follow the process from Homework 7. Check the Homework 7 solutions for an example of how to clean the data based on reasonable values for the `TOT_ENROLL` variable.

In [2]:

```
df = pd.read_csv('colleges.csv')
df = df[["FT_ENROLL", "PT_ENROLL", "TOT_ENROLL", "TOT_EMPLOY"]]
df = df[df["TOT_ENROLL"] > 0]
df["majority_pt"] = df["PT_ENROLL"] > df["FT_ENROLL"]
```

In [3]:

```
df.head()
```

Out[3]:

| | FT_ENROLL | PT_ENROLL | TOT_ENROLL | TOT_EMPLOY | majority_pt |
|---|---|---|---|---|---|
| **0** | 1226 | 440 | 1666 | 356 | False |
| **1** | 2075 | 2141 | 4216 | 697 | True |
| **2** | 3475 | 916 | 4391 | 5574 | False |
| **3** | 1378 | 130 | 1508 | 319 | False |
| **5** | 333 | 3074 | 3407 | 91 | True |

In [4]:

```
df.shape
```

Out[4]:

```
(128, 5)
```

## 2. What Population Does a College Serve? [10 points]

The University of Illinois reports that approximately 15% of students enrolled part-time.

A university diversity officer would like to see if that 15% is also a reasonable value for a different characteristic of colleges and universities: the proportion of colleges that serve a majority of part-time students (as opposed to full-time students).

**a)** Write out our hypotheses to test in this question. Be sure to use appropriate notation, and define our parameter of interest.

$H_0 : p = 0.15$ vs. $H_a : p \neq 0.15$

Here, p is the population proportion of all colleges and universities that serve a majority of part-time students (as opposed to full-time students).

**b)** Check whether the conditions are met for our hypothesis test to be valid. (No need to state each condition, but you may if it helps you.)

**Note:** you may continue your analysis as if these conditions were met, even if they are not.

1. random sample [check, from background]
2. n is less than 10% of the population [check, as n = 128, and there are more than 1280 colleges or universities]
3. $np_0$ and $n(1 - p_0)$ are both at least 10 [check, see below for the calculation]

In [5]:

```
n = 128
p_0 = 0.15
```

```
print(n*p_0)
print(n*(1-p_0))
```

```
19.2
108.8
```

**c)** Calculate the test statistic and corresponding *p*-value for this test.

In [6]:

```
phat = df["majority_pt"].mean()
print(phat)
```

```
0.203125
```

In [7]:

```
test_stat = (phat - p_0) / ((p_0 * (1 - p_0) / n) ** (1/2))
pval = 2 * (1 - norm.cdf(test_stat))
print("My test statistic is: ", test_stat)
print("My p-value is: ", pval)
```

```
My test statistic is:  1.6832508230603465
My p-value is:  0.09232654595631029
```

I can calculate my test statistic according to the following formula: $z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$. Because this is a two-sided hypothesis test, I want to shade both tails. To do this, I first want to find the area of one tail (done with 1 - cdf because my test statistic is positive and the cdf finds the area to the left of my value), and then multiply by 2 in order to find the area of both tails. The resulting value is my p-value.

**d)** Based on your hypothesis test results, what **decision** would you make about our hypotheses at a significance level of $0.10$?

My p-value is just barely smaller than my $\alpha$ (0.0923 < 0.10), so I would reject the null hypothesis.

**e)** Which of the following statements is correct about what our significance level represents?

- Statement 1: The significance level is the probability that the null hypothesis is true.
- Statement 2: The significance level is the probability of getting a test statistic as extreme or more extreme than observed, assuming the null hypothesis is true.
- Statement 3: The significance level is the proportion of experiments that we would reject the null hypothesis when the null hypothesis is true.
- Statement 4: The significance level is the proportion of times that the alternative hypothesis is true.

**Replace the X with a number below.**

**Statement 3 is correct.**

Statement 1 is incorrect. In general, the probability that the null hypothesis is true is either 0 or 1, as the null hypothesis is either true or it isn't.

Statement 2 is the definition of the p-value.

Statement 3 is a definition of the significance level. The significance level represents the cutoff point at which you reject the null hypothesis, so it is the probability of rejecting the null hypothesis when the null hypothesis is true. The proportion here is simply another way to discuss a probability.

Statement 4 is incorrect, with an argument similar to Statement 1.

# 3. College and University Employees [12 points]

Because many colleges are small, I believe that the average number of employees of all post-secondary educational facilities (schools) will be less than 481 people. I want to make a decision about this claim using data. Help me to perform an appropriate hypothesis test to test my theories.

For this question, I'll use a significance level of $\alpha = 0.01$.

**a)** Write out your hypotheses. Be sure to use appropriate notation and to define the parameter of interest.

$H_0 : \mu = 481$

$H_a : \mu < 481$

where $\mu$ is the population mean number of employees for all post-secondary educational facilities.

**b)** State & check the necessary assumptions for the hypothesis test to be valid.

State:

1. Our observations are independent, which happens if:
   - the sample is a random sample, and
   - the sample size is less than 10% of the population size
2. Our sampling distribution is Normal, which happens if:
   - the population is Normally distributed, or
   - the sample size is large enough such that the CLT applies, or n > 30

Check:

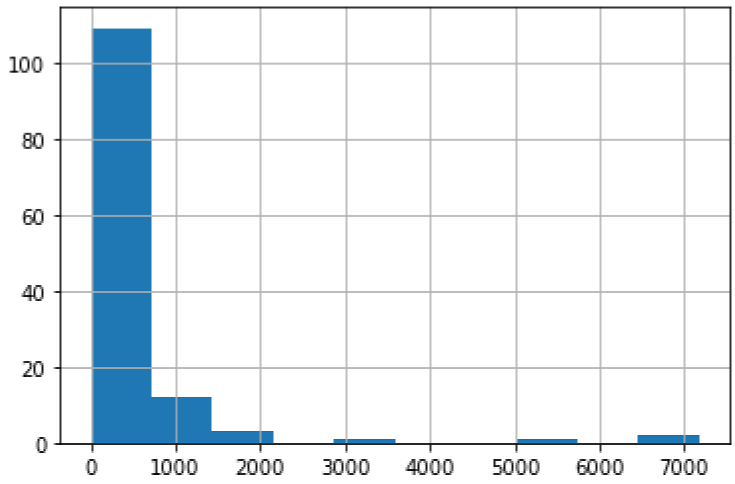
1. random sample [check, from the background]
2. sample size is less than 10% of the population size [check, there are likely more than 1280 post-secondary schools in the US]
3. the population is normally distributed [not reasonable, based on the histogram below. We would expect a symmetric distribution from the sample if the population were normally distributed] **BUT**
4. n = 128 > 30 [check, and since this condition is met, then our normality condition succeeds even though condition 3 above failed]

In [8]:

```
df['TOT_EMPLOY'].hist()
```

Out[8]:

```
<AxesSubplot:>
```



**c)** Calculate the test statistic and corresponding *p*-value based on the provided sample of colleges.

In [9]:

```
xbar = df['TOT_EMPLOY'].mean()
s = df['TOT_EMPLOY'].std()
n = df.shape[0]
print('sample mean:', xbar)
print('sample standard deviation:', s)
print('sample size:', n)
```

```
sample mean: 438.2890625
sample standard deviation: 1055.30537545773
sample size: 128
```

In [10]:

```
test_stat = (xbar - 481) / (s / n ** (1/2))
pval = t.cdf(test_stat, n - 1)
print("My test statistic is: ", test_stat)
print("My p-value is: ", pval)
```

```
My test statistic is:  -0.4578950395128658
My p-value is:  0.3239051588757881
```

I can calculate my test statistic according to the following formula: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.

Because this is a one-sided hypothesis test, I want to shade just the test statistics in the direction of my alternative hypothesis (less than). To do this, I want to find the area to the left of my test statistic (done with cdf). I will use the t distribution with $n-1=127$ degrees of freedom to find this value, since I also needed to estimate $\sigma$ with $s$ when I calculated by test statistic.

**d)** Based on the evidence calculated so far, what can you say about our theory of interest? Be sure to give a complete conclusion in the context of the problem.

My *p*-value of $0.324 > 0.01 = \alpha$, so I will fail to reject my null hypothesis.

While my sample mean is in fact less than 481 (which is consistent with the alternative hypothesis), this is not enough evidence to convince the skeptic who believes that the null hypothesis is true.

My complete conclusion is: I do not have sufficient evidence to suggest that the population mean number of employees for all post-secondary schools is less than 481 people.

**e)** Interpret the *p*-value. That is, explain what the specific *p*-value that you calculated means.

My *p*-value is the amount of evidence I have against the null hypothesis. This is how much evidence I have that will convince the skeptic who believes the null hypothesis is true.

The specific, formal interpretation of the *p*-value is: If the population mean number of employees for the population of all post-secondary schools really is 481 (the null hypothesis is true), I would expect to have a sample of 128 schools with a sample mean of what I observed from this sample (438) or less in about 32.4% of my possible samples.

Equivalently, I could write my interpretation in terms of test statistics. The probability of obtaining a test statistic of t = -0.458 or less is 32.4%, assuming that the true population mean number of employees for post-secondary schools really is 481.

My skeptic would say that this is too common of a result and isn't enough evidence to meet the threshold of evidence needed (1%, in this case), to adjust their beliefs.

---

# Case Study 2: U.S. County Unemployment Rate and Metropolitan Areas

In this case study, we will explore the statistical concepts that we've learned this week while also exploring the relationship between the **unemployment rate** of U.S. counties and whether the county is a **metropolitan area** or not. In this analysis, our cleaned U.S. counties dataset will serve as our population of all U.S. counties. Other than those counties that have been removed, this dataset is in fact the population of U.S. coutnies. While we usually do not have the whole population at our disposal when we conduct inference, we will use this population to "check" our answers so we can gain a deeper understanding of what is going on "behind the scenes" when we conduct inference on a population parameter.

## 4. Read and prepare the data [3 points]

We will be analyzing just the **unemployment_rate** and **metro** variables from the **county.csv** dataframe.

1. Read the county.csv into a dataframe, call it df, and display the first five rows.
   - *This dataset has missing values! Specifically, the phrase 'data unavailable' represents missing values in this csv. Make sure you encode these values as NaN when you read the data in.*
2. Create a pandas dataframe that is comprised of just the **unemployment_rate** and the **metro** columns. Then drop all rows in this dataframe that have missing values.

In [11]:

```
df=pd.read_csv('county.csv', na_values=['data unavailable'])
df.head()
```

Out[11]:

| | name | state | pop2000 | pop2010 | pop2017 | pop_change | poverty | homeownership | multi_unit | unem |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Autauga County | Alabama | 43671.0 | 54571 | 55504.0 | 1.48 | 13.7 | 77.5 | 7.2 | |
| **1** | Baldwin County | Alabama | 140415.0 | 182265 | 212628.0 | 9.19 | 11.8 | 76.7 | 22.6 | |
| **2** | Barbour County | Alabama | 29038.0 | 27457 | 25270.0 | -6.22 | 27.2 | 68.0 | 11.1 | |
| **3** | Bibb County | Alabama | 20826.0 | 22915 | 22668.0 | 0.73 | 15.2 | 82.9 | 6.6 | |
| **4** | Blount County | Alabama | 51024.0 | 57322 | 58013.0 | 0.68 | 15.6 | 82.0 | 3.7 | |

In [12]:

```
df=df[['unemployment_rate','metro']].dropna()
df
```

Out[12]:

| | unemployment_rate | metro |
|---|---|---|
| **0** | 3.86 | yes |
| **1** | 3.99 | yes |
| **2** | 5.90 | no |
| **3** | 4.39 | yes |
| **4** | 4.02 | yes |
| **...** | ... | ... |

|  | unemployment_rate | metro |
|---|---|---|
| **3137** | 4.55 | no |
| **3138** | 2.99 | no |
| **3139** | 4.50 | no |
| **3140** | 4.08 | no |
| **3141** | 3.98 | no |

3139 rows × 2 columns

# 5. Parameter Information [8 points]

In this analysis, we will consider two populations:

- the population of counties that are metropolitan areas, and
- the population of counties that are not metropolitan areas

**a)** First, create two dataframes:

- one that is comprised of the unemployment rates of all metropolitan counties, and
- one that is comprised of the unemployment rates of all non-metropolitan counties.

In [13]:

```
df_metro=df[df['metro']=='yes']
df_metro.head()
```

Out[13]:

|  | unemployment_rate | metro |
|---|---|---|
| **0** | 3.86 | yes |
| **1** | 3.99 | yes |
| **3** | 4.39 | yes |
| **4** | 4.02 | yes |
| **7** | 4.93 | yes |

In [14]:

```
df_metro.shape[0]
```

Out[14]:

```
1165
```

In [15]:

```
df_nonmetro=df[df['metro']=='no']
df_nonmetro.head()
```

Out[15]:

|  | unemployment_rate | metro |
|---|---|---|
| **2** | 5.90 | no |
| **5** | 4.93 | no |
| **6** | 5.49 | no |
| **8** | 4.08 | no |
| **9** | 4.05 | no |

In [16]:

```
df_nonmetro.shape[0]
```

Out[16]:

```
1974
```

**b)** Next, we will calculate two relevant parameters for each population. For each of these two populations, calculate the population mean unemployment rate and the population standard deviation unemployment rate.

In [17]:

```
pop_mean_metro=df_metro['unemployment_rate'].mean()
pop_mean_metro
```

Out[17]:

```
4.397957081545066
```

In [18]:

```
pop_std_metro=df_metro['unemployment_rate'].std()
pop_std_metro
```

Out[18]:

```
1.2991800671784013
```

In [19]:

```
pop_mean_nonmetro=df_nonmetro['unemployment_rate'].mean()
pop_mean_nonmetro
```

Out[19]:

```
4.737436676798381
```

In [20]:

```
pop_std_nonmetro=df_nonmetro['unemployment_rate'].std()
pop_std_nonmetro
```

Out[20]:

```
1.8124518106510843
```

**c)** What is the difference between the population mean unemployment rate of metropolitan counties and the population mean unemployment rate of non-metropolitan counties (ie. $\mu_{metro} - \mu_{nonmetro}$)?

Is this a statistic or a parameter?

In [21]:

```
pop_mean_metro-pop_mean_nonmetro
```

Out[21]:

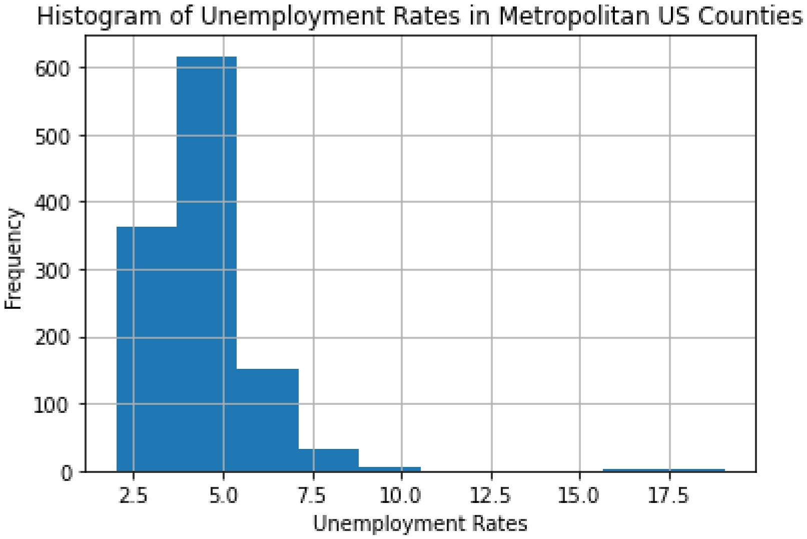
```
-0.33947959525331495
```

Answer to statistic or parameter: **parameter**

**d)** Plot a histogram of distribution of unemployment rates for each of the two types of counties (whether in a metropolitan area or not). Describe the **shape** for each distribution.

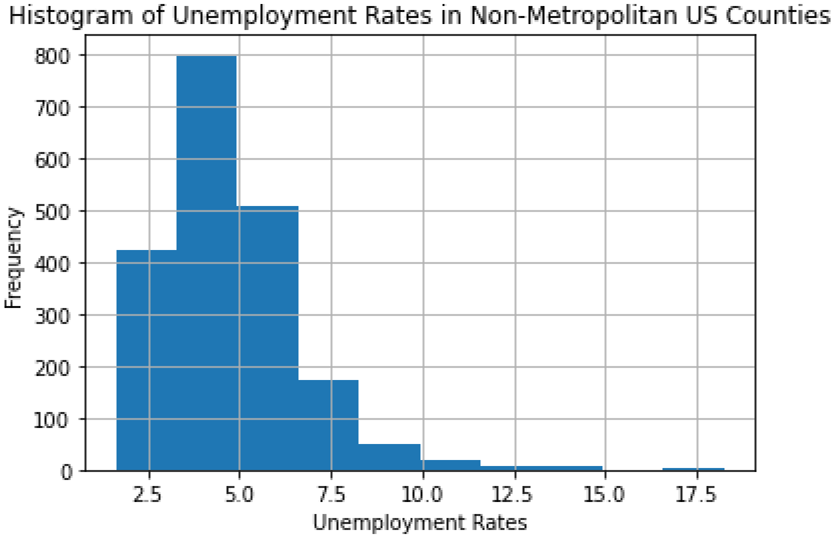
In [22]:

```
df_metro['unemployment_rate'].hist()
plt.xlabel('Unemployment Rates')
plt.ylabel('Frequency')
plt.title('Histogram of Unemployment Rates in Metropolitan US Counties')
plt.show()
```



In [23]:

```
df_nonmetro['unemployment_rate'].hist()
plt.xlabel('Unemployment Rates')
plt.ylabel('Frequency')
plt.title('Histogram of Unemployment Rates in Non-Metropolitan US Counties')
plt.show()
```



Both of these graphs are unimodal and right skewed (with a long right tail).

Since neither graph is very symmetric, neither population can be normally distributed. Because these are population distributions, I can definitively say that these populations are not normal. Here, I don't need to discuss how a sample distribution mirrors the shape of the population distribution.

## 6. Sampling Distribution for Difference of Means [14 points]

First, *suppose* we were to create a sampling distribution of sample mean unemployment rate differences (ie. a distribution of values of $\bar{x}_{metro} - \bar{x}_{nonmetro}$, where $\bar{x}_{metro}$ is the mean of a random sample of $n_1 = 50$ metropolitan counties and $\bar{x}_{nonmetro}$ is the mean of a random sample of $n_2 = 50$ non-metropolitan counties).

**a)** What would you *expect* the mean and standard deviation of this sampling distribution to be?

In [24]:

```
sampling_dist_mean=pop_mean_metro-pop_mean_nonmetro
sampling_dist_mean
```

```
Out[24]:  -0.33947959525331495
```

```
In [25]:  sampling_dist_std=np.sqrt((pop_std_metro**2)/50 + (pop_std_nonmetro**2)/50)
          sampling_dist_std
```

```
Out[25]:  0.3153680520561989
```

$$E[\bar{X}_{metro} - \bar{X}_{nonmetro}] = \mu_{metro} - \mu_{nonmetro} = 4.398 - 4.737 = -0.339$$

and

$$SD[\bar{X}_{metro} - \bar{X}_{nonmetro}] = \sqrt{\frac{\sigma^2_{metro}}{n_{metro}} + \frac{\sigma^2_{nonmetro}}{n_{nonmetro}}} = \sqrt{\frac{1.299^2}{50} + \frac{1.812^2}{50}} = 0.340$$

**b)** Would this sampling distribution of differences of sample means be approximately normal? Explain.

Yes. Because the Central Limit Theorem conditions for sample mean differences hold, this sampling distribution is approximately normal.

1. Sample of metropolitan counties is randomly selected from the population of metropolitan counties.
2. $n_{metro} = 50 < 10\%$ of the population of metropolitan counties (1165)
3. $n_{metro} = 50 > 30$ OR ~~the population distribution of unemployment rates of all metropolitan counties is normal~~
4. Sample of nonmetropolitan counties is randomly selected from the population of metropolitan counties.
5. $n_{nonmetro} = 50 < 10\%$ of the population of nonmetropolitan counties (1974)
6. $n_{nonmetro} = 50 > 30$ OR ~~the population distribution of unemployment rates of all nonmetropolitan counties is normal~~
7. The two samples of counties are independent from each other. *(We know this because there is not a pairwise relationship in the way in which the two samples of counties were not selected.)*

*More information:*

- We can infer that the population distribution of unemployment rates of all metropolitan counties is not normal, becuase the sample distribution of unemployment rates of metropolitan counties does not have the shape of a normal distribution and these two distributions tend to mirror each other.
- We can infer that the population distribution of unemployment rates of all nonmetropolitan counties is not normal, becuase the sample distribution of unemployment rates of nonmetropolitan counties does not have the shape of a normal distribution and these two distributions tend to mirror each other.

**c)** Now, we would check your *expectations* in parts **a** and **b** above by *actually* creating a sampling distribution of differences of sample mean unemployment rates (i.e. a distribution of values of $\bar{x}_{metro} - \bar{x}_{nonmetro}$, where $\bar{x}_{metro}$ is the mean of a random sample of $n_1 = 50$ metropolitan counties and $\bar{x}_{nonmetro}$ is the mean of a random sample of $n_2 = 50$ non-metropolitan counties).

In the space below, create this sampling distribution with 1000 values of $\bar{x}_{metro} - \bar{x}_{nonmetro}$.

First let's create a sampling distribution of sample mean unemployment rates (from the population of metropolitan counties, where each of our samples are of size $n_{metro} = 50$.

```
In [26]:  sampling_dist_diff=[]

          for i in range(1000):
              # First we collect a random sample from the metro population
              sample_metro=df_metro.sample(50)
```

```
    # Next we collect a random sample from the non-metro population
    sample_nonmetro = df_nonmetro.sample(50)

    # We can calculate our statistic of interest
    xbar1 = sample_metro['unemployment_rate'].mean()
    xbar2 = sample_nonmetro['unemployment_rate'].mean()
    xbar1_xbar2 = xbar1 - xbar2

    # Next we add the corresponding difference of sample means to the list
    sampling_dist_diff.append(xbar1_xbar2)

sampling_dist=pd.DataFrame({'diff_sample_means':sampling_dist_diff})
sampling_dist
```

Out[26]:

| | diff_sample_means |
|---|---|
| **0** | -0.4170 |
| **1** | -0.2346 |
| **2** | -0.1798 |
| **3** | -0.8464 |
| **4** | -0.3882 |
| **...** | ... |
| **995** | -0.5074 |
| **996** | -0.9308 |
| **997** | -0.3180 |
| **998** | -0.6474 |
| **999** | -0.3796 |

1000 rows × 1 columns

**d)** Calculate the mean of the simulated sampling distribution we just created. This is an *estimate* of the mean for the sampling distribution. Compare this to the *theoretical* sampling distribution mean that you calculated in part **a**.

In [27]:

```
sampling_dist.mean()
```

Out[27]:

```
diff_sample_means   -0.343994
dtype: float64
```

The estimated mean (calculated from the simulated sampling distribution) is -0.34. Compared to the theoretical mean from part **a** of -0.339, I find that the two values are fairly similar to each other.

**e)** Calculate the standard deviation of the simulated sampling distribution we just created. This is an *estimate* of the standard deviation (or standard error) for the sampling distribution. Compare this to the *theoretical* sampling distribution standard deviation that you calculated in part **a**.

In [28]:

```
sampling_dist.std()
```

Out[28]:

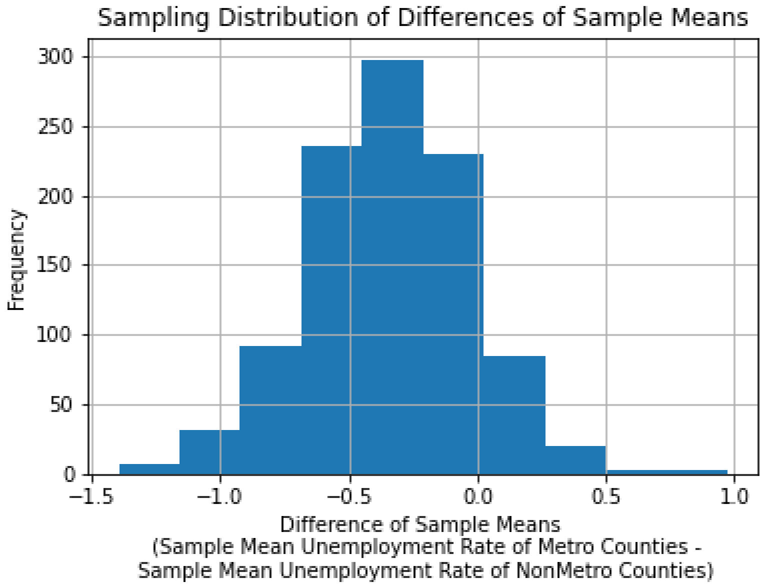
```
diff_sample_means    0.312221
dtype: float64
```

The estimated standard deviation (calculated from the simulated sampling distribution) is 0.323. Compared to the theoretical standard error from part **a** of 0.315, I again see that these two values are very similar to each other.

**f)** Finally, plot a histogram of this sampling distribution that you just created. Make sure that you appropriately label the x-axis, the y-axis, and the title of this plot. Describe the shape.

In [29]:

```
sampling_dist['diff_sample_means'].hist()
plt.title('Sampling Distribution of Differences of Sample Means')
plt.xlabel('Difference of Sample Means \n (Sample Mean Unemployment Rate of Metro Counties
plt.ylabel('Frequency')
plt.show()
```



Our sampling distribution does look approximately normal, as we expected.

Remember to keep all your cells and hit the save icon above periodically to checkpoint (save) your results on your local computer. Once you are satisified with your results restart the kernel and run all (Kernel -> Restart & Run All). **Make sure nothing has changed**. Checkpoint and exit (File -> Save and Checkpoint + File -> Close and Halt). Follow the instructions on the Homework 4 Canvas Assignment to submit your notebook to GitHub.