Homework 09: Small Sample Hypothesis Testing, Simple Linear Regression

---

Name: Julia Troni

This assignment is due on Canvas by **6:00PM on Friday November 11**. Your solutions to theoretical questions should be done in Markdown directly below the associated question. Your solutions to computational questions should include any specified Python code and results as well as written commentary on your conclusions. Remember that you are encouraged to discuss the problems with your classmates, but **you must write all code and solutions on your own**.

NOTES:

- Any relevant data sets should be available in the Homework 01 assignment write-up on Canvas. To make life easier on the grader if they need to run your code, do not change the relative path names here. Instead, move the files around on your computer.
- If you're not familiar with typesetting math directly into Markdown then by all means, do your work on paper first and then typeset it later. Remember that there is a reference guide linked on Canvas on writing math in Markdown. **All** of your written commentary, justifications and mathematical work should be in Markdown.
- Because you can technically evaluate notebook cells is a non-linear order, it's a good idea to do Kernel → Restart & Run All as a check before submitting your solutions. That way if we need to run your code you will know that it will work as expected.
- It is **bad form** to make your reader interpret numerical output from your code. If a question asks you to compute some value from the data you should show your code output **AND** write a summary of the results in Markdown directly below your code.
- This probably goes without saying, but... For any question that asks you to calculate something, you **must show all work and justify your answers to receive credit**. Sparse or nonexistent work will receive sparse or nonexistent credit.

Here are some imports that you might find handy:

In [1]:
```
import numpy as np
from scipy import stats
from scipy.stats import t
import pandas as pd
import matplotlib.pylab as plt
%matplotlib inline
```

Recall the general steps in hypothesis testing:

- Determine if the situation calls for a Z-test or a T-test.

- State the null hypothesis
- State the alternate hypothesis
- Set alpha
- collect data
- calculate a test statistic
- Construct acceptance/rejection regions
- Based on the test statistic and the acc./rej. regions, draw a conclusion about the null hypothesis.

Problem 1

In this question you are a quality control engineer inspecting parts made at Cube Aerospace Manufacturing. You will need to decide whether or not to stop the manufacturing process to adjust the calibration of the machines making parts.

The part being inspected at work today is for aircraft. The part has a small port (hole) that must be tightly controlled with a 0.02 dm diameter otherwise catastrophic failure could result in fuel access (too much or too little) for the aircraft.

At various times the engineer takes a small sample of the components from the production line and measures the port diameter and possibly stops the assembly line to make adjustments to the machines if needed.

At one of the these times four units are taken off the line and measured. The resulting port measurements (in dm) came in at: $0.021, 0.019, 0.023, 0.020$.

Assuming the port diameters of interest are normally distributed, determine at the 1% level of significance, if there is sufficient evidence in the sample to conclude that processing stop since an adjustment is likely needed.

Part A

(2 points) Is this a Z-test or a T-test? Describe what you know about the test and its distribution.

Solution:

This is a T-test becasue we have a sample size n=4 <30 which is normally distributed, however we do not know the population variance/standard deviation so we must estimate it with the sample standard deviation.

We know $n = 4, \mu = 0.02, \alpha = 0.01, \bar{X} = 0.02075, s = 0.001707$ (work shown in part C)

Part B

(2 points) What is the null hypothesis and what is the alternate hypothesis?

Solution:

- Null hypothesis: $H_0 : \mu = 0.02$
- Alternate hypothesis: $H_a : \mu \neq 0.02$

Part C

(3 points) Calculate the proper test statistic.

In [2]:
```
#the test statistic is the sample mean Xbar
sample= np.array([0.021,0.019,0.023,0.020])
Xbar=np.mean(sample)
print("The test statistic is the sample mean Xbar= ", Xbar)

#we will also need the sample standard deviation for the T-test so I calculated it here
s=np.std(sample, ddof=1)
```

```
The test statistic is the sample mean Xbar=  0.02075
```

Part D

(3 points) What is/are the critical value(s)?

Solution:

- The t critical value for a two tailed test can be found using $stats.t.ppf(1-\frac{\alpha}{2}, n-1)$ thus $stats.t.ppf(1-\frac{0.01}{2}, 4-1) = 5.8409$

  So the critical values are 5.8409 and -5.8409

- Then, we can calculate the tscore $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{0.02075-0.02}{\frac{0.00170}{\sqrt{4}}} = 0.878310065.$

  Thus since $0.8783 < 5.8409$ (i.e. tscore < critical value) we fail to reject the null hypothesis at the 99% confidence level

In [3]:
```
#xbar=0.02075
#mu= 0.02
#s=0.00170
#n=4
#alpha= 0.01

#tcriticalvalue from stats.t.ppf
tcriticalvalue= stats.t.ppf(1-0.01/2,4-1)
print("The tcritical value is ", tcriticalvalue)

CI= [Xbar-(tcriticalvalue*s/2), Xbar+(tcriticalvalue*s/2)] #[0.02, 0.0215]
CI, Xbar ## in CI so accept

print ("The 99% confidence interval is ", CI)
print("Since the test statistic Xbar= 0.02075 lies within this interval, we fail to rej
```

```
The tcritical value is  5.84090929975643
The 99% confidence interval is  [0.015762374164746695, 0.025737625835253308]
Since the test statistic Xbar= 0.02075 lies within this interval, we fail to reject our
null hypothesis
```

Part E

(2 points) What is the conclusion to our hypothesis test and what does it mean with respect to this problem?

Solution:

- The t critical values are 5.8409 and -5.8409
- Then, tscore $t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{0.02075-0.02}{\frac{0.00170}{\sqrt{4}}} = 0.878310065.$

  Thus since $0.8783 < 5.8409$ (i.e. tscore < critical value) we fail to reject the null hypothesis at the 99% confidence level

Alternatively, 99% confidence interval is [0.015762374164746695, 0.025737625835253308] Since the test statistic Xbar= 0.02075 so the 99% confidence interval does contain the test statistic, so we fail to reject our null hypothesis. Thus, at the 1% significane level we have sufficent evidence to conclude that the process does NOT need stop since an adjustment is NOT needed.

Part F

(3 points) Demonstrate how you would come to this same conclusion using the p-value approach.

Some documentation for `stats.ttest_1samp` :
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html

and documentation for `t.cdf` :
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.t.html

In [4]:
```
# Code your solution here using stats.ttest_1samp

t_test_statistic, pvalue= stats.ttest_1samp(sample, popmean=0.02)
print("The t value of the test statistic is ", t_test_statistic)
print("The corresponding p value is " , pvalue)
```

```
The t value of the test statistic is  0.8783100656536806
The corresponding p value is  0.4444380851347304
```

In [5]:
```
# using t.cdf()

t= (Xbar-0.02)/(s/np.sqrt(4))
pval= 2*stats.t.cdf(-t,4-1) #0.4444
print("pvalue " , pval)
print("t test statistic ", t)
```

```
pvalue  0.4444380851347304
t test statistic  0.8783100656536806
```

$t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} = \frac{0.02075-0.02}{\frac{0.00170}{\sqrt{4}}} = 0.878310065$

$pvalue = 2 \cdot t_{\alpha/2,v} = \texttt{2*stats.t.cdf(-t,n-1)} = 0.4444$

Then since $pvalue = 0.4444 > \alpha = 0.01$ we DO NOT reject the null hypothesis and conclude that there is sufficient evidence to believe, at the 1% significance level, that production process can continue without stopping to adjust the machines

Review the code below which graphs PDF curves and CDF curves and...

Part G

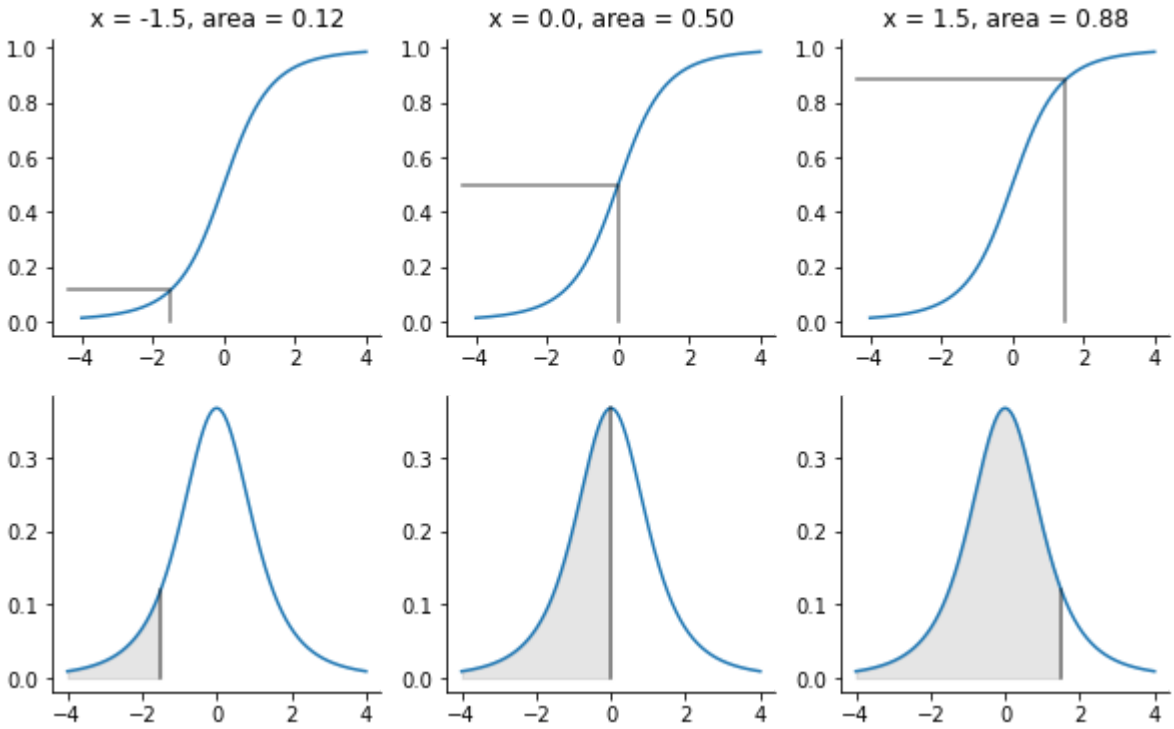
(2 points) ...fill in the requested questions/comments found in the code below.

In [6]:

```
t_dist = stats.t(3)
# What does stats.t(3) mean?
# hint: https://docs.scipy.org/doc/scipy-0.13.0/reference/generated/scipy.stats.t.html
# ANSWER: stats.t(3)   is a Student's T continuous random variable object with many met

t_values = np.linspace(-4, 4, 1000)
# What is contained in the variable 't_values'?
# ANSWER: t_values contains 1000 values equally spaced between -4 and 4

#########################################################

   # Set 1 of t-values.
#Lt = -5.84
#Mt = 0
#Ut = 5.84

   # Set 2 of t-values.
Lt = -1.5
Mt = 0
Ut = 1.5

# Try the following code with both sets of t-values above,
#   one set at a time.
# Of course you will need to comment one set out and
#  un-comment the other set when you try each set.

example_values = (Lt, Mt, Ut)
pdf_values = t_dist.pdf(t_values)
cdf_values = t_dist.cdf(t_values)
fill_color = (0, 0, 0, 0.1)  # Light gray in RGBA format.
line_color = (0, 0, 0, 0.5)  # Medium gray in RGBA format.
fig, axes = plt.subplots(2, len(example_values), figsize=(10, 6))
for i, x in enumerate(example_values):
    cdf_ax, pdf_ax = axes[:, i]
    cdf_ax.plot(t_values, cdf_values)
    pdf_ax.plot(t_values, pdf_values)

    # Fill area at and to the left of x.
    pdf_ax.fill_between(t_values, pdf_values,
                        where=t_values <= x,
                        color=fill_color)
```

```
    # Probability density at this value.
    pd = t_dist.pdf(x)

    # Line showing position of x on x-axis of PDF plot.
    pdf_ax.plot([x, x],
                [0, pd], color=line_color)

    # Cumulative distribution value for this x.
    cd = t_dist.cdf(x)

    # Lines showing x and CDF value on CDF plot.
       # x position of y axis on plot.
    x_ax_min = cdf_ax.axis()[0]
    cdf_ax.plot([x, x, x_ax_min],
                [0, cd, cd], color=line_color)
    cdf_ax.set_title('x = {:.1f}, area = {:.2f}'.format(x, cd))

    # Hide top and right axis lines and ticks to reduce clutter.
    for ax in (cdf_ax, pdf_ax):
        ax.spines['right'].set_visible(False)
        ax.spines['top'].set_visible(False)
        ax.yaxis.set_ticks_position('left')
        ax.xaxis.set_ticks_position('bottom')

# Area of PDF at and to the left of 1.5
t_dist.cdf(Ut)
```

Out[6]: 0.8847080673775886



Part H

(2 points) What do these series of graphs represent?

Solution:

Set one shows the 1% and 99% confidence intervals for a t distribution Set two shows the 12% and 88% confidence intervals for a t distribution

Problem 2

Supply line issues have caused a boom in the sale of used cars. In this question you are advising a start-up called CU.com (Cars Used .com). CU.com would like to know appropriate prices for used cars.

You decide to sample some local car dealerships and you find the following data:

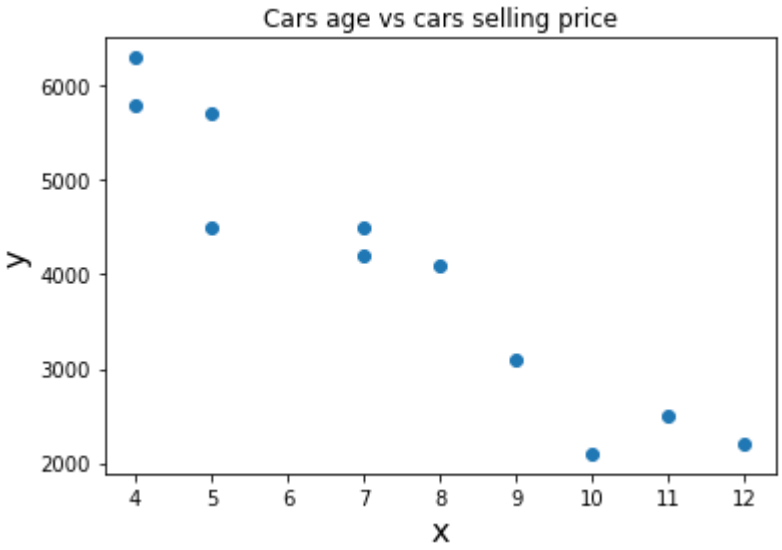
Example:

Cars Age	Cars Price
(in years)	(in dollars)
4	6300
4	5800
5	5700
5	4500
7	4500
7	4200
8	4100
9	3100
10	2100
11	2500
12	2200

Part A

(3 points) Make a scatter plot of this data to determine if there is a relationship between the cars age and the cars selling price.

In [7]:
```
# Code your solution here:

x = [4,4,5,5,7,7,8,9,10,11,12]
y = [6300,5800,5700,4500,4500,4200,4100,3100,2100,2500,2200]

fig, ax = plt.subplots()
line = ax.scatter(x,y)
ax.set_xlabel("x", fontsize=16)
ax.set_ylabel("y", fontsize=16)
ax.set_title("Cars age vs cars selling price")
plt.show()
```



Part B

(2 points) After viewing the scatterplot, how would you describe the relationship?

Solution:

There is a negative relationship between the cars age and the cars selling price. As car age (predictor) increases, the selling price (response) decreases. Likewise, as car age (predictor) decreases, the selling price (response) increases

Part C

(4 points) What is the regression equation for this example? i.e. What is the line of best fit?

Use TeX to write the equation, with the appropriate values, in the cell below:

In [8]:

```
X = [4,4,5,5,7,7,8,9,10,11,12]
Y = [6300,5800,5700,4500,4500,4200,4100,3100,2100,2500,2200]

#Mean of variables the each data set
mean_x = np.mean(X)
mean_y = np.mean(Y)

#Total number of data points
n = len(X)

num = 0 #calculating ∑(x-x¯)·(y-y¯)
denom = 0 #calculating ∑(x-x¯)2

for i in range(n):
    #calculating ∑(x-x¯)·(y-y¯)
    num += (X[i] - mean_x) * (Y[i] - mean_y)
    #calculating ∑(x-x¯)2
    denom += (X[i] - mean_x) ** 2

#beta is the slope
beta = np.array(num) / np.array(denom)
```

```
#alpha is y intercept
alpha= mean_y - (beta * mean_x)

print("The slope beta is ", beta)
print("The y intercept alpha is ", alpha)
print("∑(x-x¯)·(y-y¯) = ", num)
print(" ∑(x-x¯)2 = ", denom)
```

```
The slope beta is  -502.42494226327943
The y intercept alpha is  7836.258660508083
∑(x-x¯)·(y-y¯) =  -39554.545454545456
 ∑(x-x¯)2 =  78.72727272727273
```

Solution:

Note: I used the above code to help with the summations, but all relevant formulas and equations are in LaTex below, the code above is not required to understand my calculations

$\bar{x} = \frac{\sum(x)}{n} = \frac{82}{11} = 7.45$

and $\bar{y} = \frac{\sum(y)}{n} = \frac{45000}{11} = 4090.90$

Thus, $\sum(x - \bar{x}) \cdot (y - \bar{y}) = -39554.545$ and

$\sum(x - \bar{x})^2 = 78.7272$

The slope is beta,
$\beta = \frac{\sum(x-\bar{x})\cdot(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{-39554.54}{78.72} = 7836.258$

Alpha is the y intercept, $\alpha = \bar{y} - \beta\bar{x} = 4090.90 - (7.45)(7836.258) = -502.425$

Since the line of best fit is $\hat{y} = \alpha + \beta x$

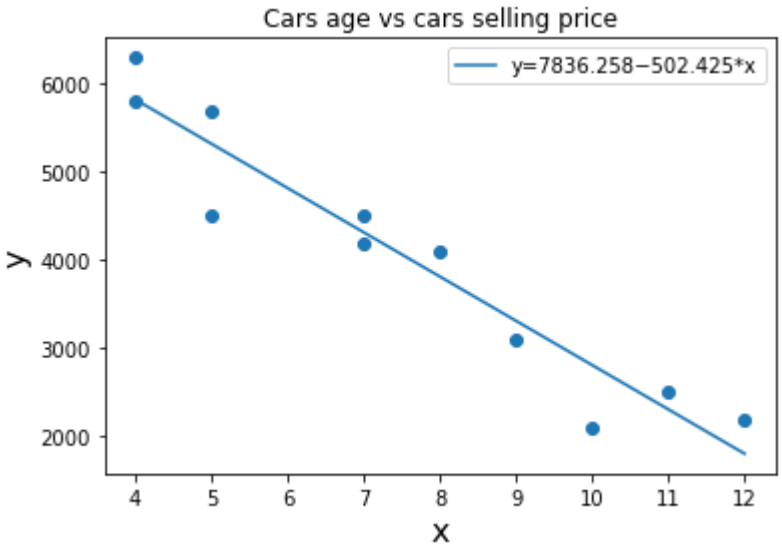
Thus, $\hat{y} = 7836.258 - 502.425 \cdot x$

Part D

(2 points) Draw the same scatterplot as above, but this time add the line of best fit on top of the scatterplot.

In [9]:

```
x = [4,4,5,5,7,7,8,9,10,11,12]
y = [6300,5800,5700,4500,4500,4200,4100,3100,2100,2500,2200]
linex = x
liney = np.array(alpha + beta*np.array(x))

fig, ax = plt.subplots()
ax.scatter(x,y)
ax.plot(linex,liney,label="y=7836.258-502.425*x")
ax.set_xlabel("x", fontsize=16)
ax.set_ylabel("y", fontsize=16)
ax.set_title("Cars age vs cars selling price")
plt.legend()
plt.show()
```



Part E

(2 points) Interpret meaning of the regression line. What does b_1 (aka β) indicate relative to this problem?

Solution:

The regression line, $\hat{y} = 7836.258 - 502.425 \cdot x$

This indicates that as the age of a car gets 1 year older, the selling price decreases by $502.425.

And from $\beta = 7836.258$ we can predict that a brand new car (0 years old) costs $7836.25

Problem 3

You have invented a new skateboard truck! You go on the TV show "Shark Tank" and Mark Cuban gives you funding for your venture.

In an attempt to market the truck properly you consider two experimental packaging designs; Design A and Design B.

Design A is sent to 11 stores and their average sales the first month is 52 units with sample standard deviation 12 units.

Design B is sent to 6 stores and their average sales the first month are 46 units with sample standard deviation 10 units.

Part A

(2 points) What is a point estimate for the difference in average sales between the two package designs and what does the point estimate mean?

Solution:

The point estimate $=\bar{x_A} - \bar{x_B} = 52 - 46 = 6$

In words, we estimate that the average sales for Design A is 6 units higher than it is for Design B.

But how accurate is this point estimate? We can answer this by creating a 95% confidence interval

for the point estimate. Follow the steps below:

Part B

(3 points) What is the critical t-value?

Solution:

In [10]:
```
# Code here if needed:
stats.t.ppf(q=1-0.05/2, df=11+6-2)
```

Out[10]:
```
2.131449545559323
```

- This is a two tailed test, so the t critical value found using $stats.t.ppf(1-\frac{\alpha}{2}, df)$ where $df = n_A - 1 + n_B - 1 = n_A + n_B - 2$

Thus the critical t value is = $stats.t.ppf(q = 1 - 0.05/2, df = 11 + 6 - 2) = 2.1314$

So the critical values are 2.1314 and -2.1314

Part C

(3 points) What is the 95% confidence interval for the point estimate? Either calculate it by 'hand' in the code or look up documentation on `stats.t.interval()`

In [11]:
```
CI= stats.t.interval(alpha=1-0.05, df=11+6-2, loc=6, scale=np.sqrt(144/11+100/6))
print("The 95% confidence interval is ", CI)
```

```
The 95% confidence interval is  (-5.627164870160184, 17.627164870160186)
```

Part D

(2 points) Interpret the CI in terms of this problem.

Solution: The CI indicates that we are 95% confident that the point estimate lies in the range (-5.627164870160184, 17.627164870160186).

This interval includes 0, meaning the mean sales per month for the 2 designs could be the same at the 95% confidence level.

Part E

Test at the 1% level of significance whether the data provide sufficient evidence to conclude that the mean sales per month of the two designs are different. Use the critical value approach.

(2 points) List the null and alternate hypothesis.

Solution:

- Null hypothesis: $H_0 : \mu_A - \mu_B = 0$
- Alternate hypothesis: $H_a : \mu_A - \mu_B \neq 0$

Part F

(2 points) What is the test statistic?

Solution:

The test statistic is $t = \frac{(\bar{x_A} - \bar{x_B}) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{(52-46)-0}{\sqrt{\frac{12^2}{11} + \frac{10^2}{6}}} = 1.099$

In [12]:
```
den=np.sqrt(144/11 + 100/6)
num=6
t=num/den
t
```

Out[12]:
```
1.0998981622920558
```

Part G

(2 points) What is the critical value?

Solution:

In [13]:
```
criticalvalue= stats.t.ppf(1-0.01/2,11+6-2)

stats.t.ppf(1-0.01/2,11+6-2)
```

Out[13]:
```
2.946712883338615
```

- This is a two tailed test, so the t critical value found using $stats.t.ppf(1 - \frac{\alpha}{2}, df)$ where $df = n_A - 1 + n_B - 1 = n_A + n_B - 2$

Thus the critical t value is = $stats.t.ppf(q = 1 - 0.01/2, df = 11 + 6 - 2) = 2.94671$

So the critical values are 2.94671 and -2.94671

Part H

(2 points) Interpret your findings with respect to this problem.

Solution:

Thus our rejection region is t > criticalvalue or t < -critical value.

So if t > 2.946 or t < -2.946, then we will reject our null hypothesis at the 99% confidence interval.

But 1.099 < 2.9467, that is t < critical value, which is NOT in the rejection region, so we FAIL TO REJECT our null hypothesis and we conclude at the 99% confidence interval that the data provides sufficient evidence that the mean sales per month are different.

So at the 1% level of significance the data provide sufficient evidence to conclude that the mean sales per month of the two designs are different.

In [ ]: