Homework 09: Small Sample Hypothesis Testing, Simple Linear Regression

---

Name: CJ Kennedy

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 $\text{Kernel} \rightarrow \text{Restart \& Run All}$ 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** $\text{write a summary of the results}$ 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:

There are 4 samples (port measurements) taken for this test, $n = 4$. In a Z-test, we would use a normal distribution and in a T-test, we use a Student T distribution. Since $n < 30$, this is a T-test. Also, we do not know the population variance.

Part B

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

Solution:

H_0: The null hypothesis is the small ports have a diameter of 0.02 dm.

H_1: The alternate hypothesis is the small port's diameter is not 0.02 dm.

Part C

(3 points) Calculate the proper test statistic.

In [2]:
```
# Code your solution here:
x = np.array([.021, .019, .023, .020])
X = np.mean(x)
s = x.std(ddof=0) # sample standard deviation
n = 4
u0 = .02
t = (X-u0)/(s/np.sqrt(n))
print("The test statistic is",np.round(t,3))
```

```
The test statistic is 1.014
```

Part D

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

Solution:

In [3]:
```
# Code any needed work here:
alpha = .01
crit = stats.t.ppf(q=1-alpha/2, df=n-1)
print("The critical values are",np.round(crit,3),"and",np.round(-crit,3))
```

```
The critical values are 5.841 and -5.841
```

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-score was not in the critical region so we fail to reject the null hypothesis. We don't have enough evidence to support the hypothesis that, on average, the small ports' diameters are not equal to 0.02 dm.

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
data = [0.021, 0.019, 0.023, 0.020]
res = stats.ttest_1samp(a=data,popmean=.02,alternative='two-sided') # (sample array, p
p1 = res.pvalue
#p1 = p1[0]*p1[1]
#p1 = p1[1]
print("The p value =",np.round(p1,4))
```

```
The p value = 0.4444
```

In [5]:
```
# Code your solution here using t.cdf()
p2 = 2*stats.t.cdf(-t,n-1)
print("The p value =",np.round(p2,4))
print("Since the p-value is greater than 0.01, we conclude that we would not reject th
```

```
The p value = 0.3852
Since the p-value is greater than 0.01, we conclude that we would not reject the nul
l.
```

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

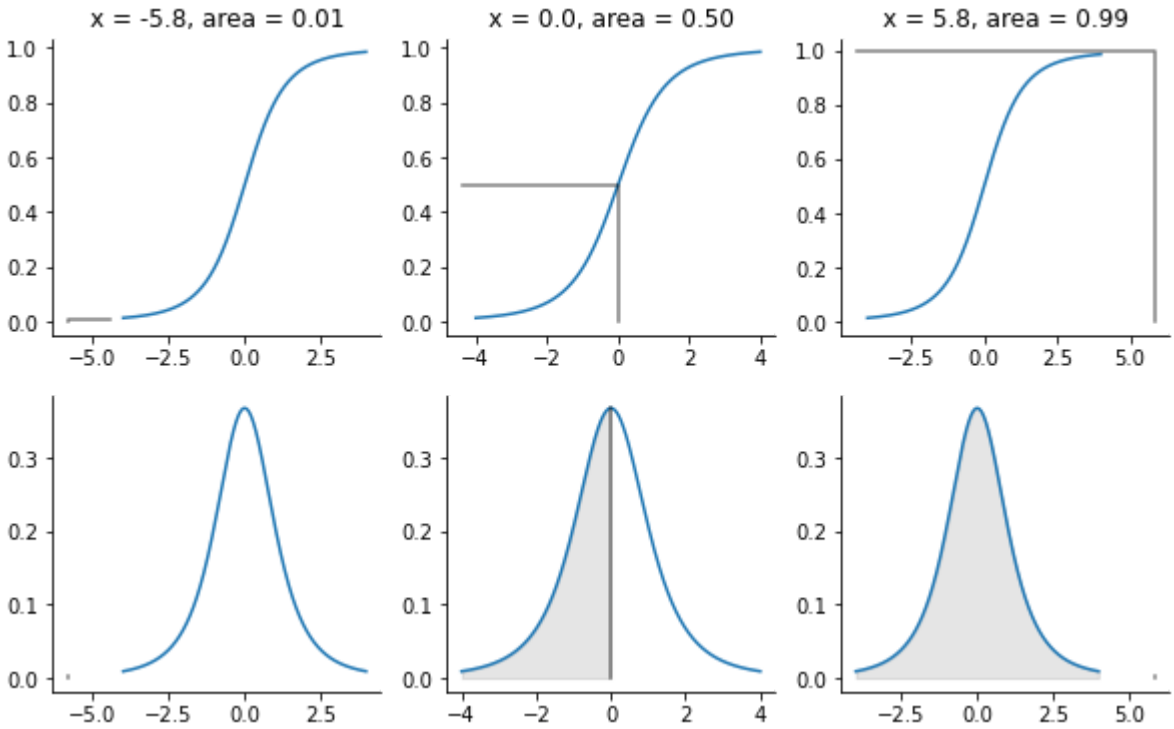
Part G

(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: a t-distrubution with 3 degrees of freedom

t_values = np.linspace(-4, 4, 1000)
# What is contained in the variable 't_values'?
# ANSWER: a linearly spaced vector from -4  to 4. 1000 total values

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

   # 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]: 0.9949978159094941



Part H

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

Solution:

There are three t-values tested for each set. These sets produce 6 graphs of the CDF and PDF. The CDF is equal to the "area" and "x" is the specific t-value.

The first three graphs are the cumulative distribution function for the t distribution.

The bottom three graphs are the probability density function for the 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:

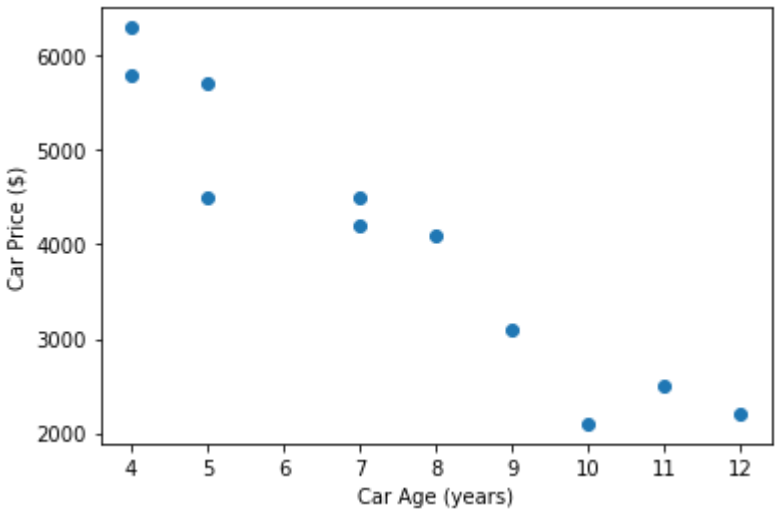
Cars Age	Cars Price
(in years)	(in dollars)
4	6300

Cars Age	Cars Price
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 = np.array([4,4,5,5,7,7,8,9,10,11,12])
y = np.array([6300,5800,5700,4500,4500,4200,4100,3100,2100,2500,2200])

plt.scatter(x, y)
plt.ylabel("Car Price ($)")
plt.xlabel("Car Age (years)")
plt.show()
```



Part B

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

Solution:

The relationship between car age and price seems to have a negative linear relationship. Specifically, as car age increases, car price decreases.

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:

Solution:

The regression equation takes the form $\hat{y} = b_0 + b_1 \cdot x$.

First, let's find the mean of x and y, $\bar{x}$ and $\bar{y}$:

$\bar{x} = \frac{4+4+5+5+7+7+8+9+10+11+12}{11} = 7.4545$

$\bar{y} = \frac{6300+5800+5700+4500+4500+4200+4100+3100+2100+2500+2200}{11} = 4090.90$

Then, we find the slope, b_1:

$b_1 = \frac{\sum(x-\bar{x})\cdot(y-\bar{y})}{\sum(x-\bar{x})}$

$\sum(x-\bar{x})^2 = 78.72$

$\sum(x-\bar{x})(y-\bar{y}) = -3.956 \cdot 10^4$

$\Rightarrow b_1 = -502.425$

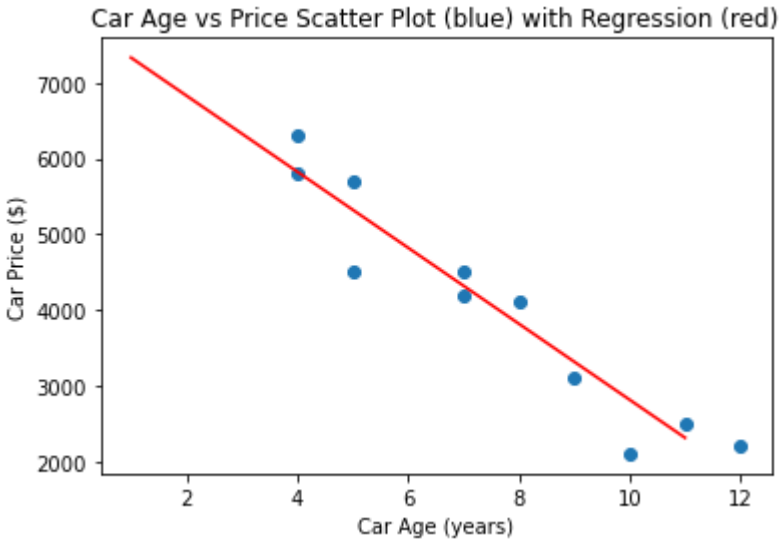
$b_0 = \bar{y} - b_1 \cdot \bar{x} = 7836.25$

$\therefore \hat{y} = 7836.25 - 502.425 \cdot x$

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 [8]:
```
# Code your solution here:
x2 = np.linspace(1,11)
y2 = 7836.25 - 502.425*x2
plt.plot(x2,y2,'r')
plt.scatter(x, y)
plt.ylabel("Car Price ($)")
plt.xlabel("Car Age (years)")
plt.title("Car Age vs Price Scatter Plot (blue) with Regression (red)")
plt.show()
```



Part E

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

Solution:

b_1 represents the slope of the regression line fit. The value is negative which shows the negative relationship. Specifically, as car age increases, the car price decreases.

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:

For Part A, let's write the known variables for Design A and Design B:

$\sigma_A = 12$

$\sigma_B = 10$

$\bar{A} = 52$

$\bar{B} = 46$

$n_A = 11$

$n_B = 6$

The point estimate for A is:

$u_A = \frac{\sigma_A}{n_A} = \frac{12}{11}$

and for B:

$u_B = \frac{\sigma_B}{n_B} = \frac{10}{6}$

The the difference would be -0.575.

Therefore, the average store sales for A is about -0.575 less than 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:

We use the below code to find the critical t-value:

In [9]:
```
# Code here if needed:
# Code your solution here:
A = 52; B = 46;
s1 = 12; s2 = 10;
n1 = 11; n2 = 6;

uDiff = -.575
#t = ((A-B)-(uDiff))/np.sqrt(s1**2/n1**2+s2**2/n2**2)
alpha = .05
crit = stats.t.ppf(q=1-alpha/2, df=n1+n2-1)
print("The critical t-value is:",np.round(crit,2))
```

```
The critical t-value is: 2.12
```

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 [10]:
```
# Code solution here:

alpha = .05
```

```
df = n1+n2-2
sp2 = ( (n1-1)*s1**2+(n2-1)*s2**2 )/ (df)
u0 = 0
test = ( (A-B)-u0 )/np.sqrt( sp2*(1/n1+1/n2) )
t = stats.t.ppf(q=1-alpha/2, df=n1+n2-1)
print("The 95% CI is between:",-np.round(t,2),"and",np.round(t,2))
print("And the test statistic is:",np.round(test,2))
```

```
The 95% CI is between: -2.12 and 2.12
And the test statistic is: 1.04
```

Part D

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

Solution:

Since the test statistic was inbetween the interval and not in the rejection region, we do not reject the null hypothesis.

We would say at the 5% level of significance, the mean sales between the two designs are not different.

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:

The null hypothesis is $H_0 : \bar{A} - \bar{B} = 0$ or that the sales aren't different between the designs.

The alternate hypothesis is $H_1 : \bar{A} - \bar{B} \neq 0$ or the sale means are significantly different.

Part F

(2 points) What is the test statistic?

Solution:

The test statistic is determined with the code below.

In [11]:
```
# Code here if needed:
alpha = .01
df = n1+n2-2
sp2 = ( (n1-1)*s1**2+(n2-1)*s2**2 )/ (df)
u0 = 0
test = ( (A-B)-u0 )/np.sqrt( sp2*(1/n1+1/n2) )
t = stats.t.ppf(q=1-alpha/2, df=n1+n2-1)
```

```
print("The 99% CI is between:",-np.round(t,2),"and",np.round(t,2))
print("And the test statistic is:",np.round(test,2))
```

```
The 99% CI is between: -2.92 and 2.92
And the test statistic is: 1.04
```

Part G

(2 points) What is the critical value?

Solution:

The critical value is determined with the code below.

In [12]:
```
# Code solution here if needed:
alpha = .01
crit = stats.t.ppf(q=1-alpha/2, df=n1+n2-1)
print("The critical value is:",np.round(crit,2))
```

```
The critical value is: 2.92
```

Part H

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

Solution:

The test statistic, 1.04, is less than the critical value, 2.92. So...

We do not have enough evidence to conclude that the difference between the mean sales of Designs A and B differ.

This conclusion is done with a significance level of 1% or we are "99% confident."

In [ ]:
```

```