Kernel: Python 3 (system-wide)

Life Sciences 40

Final Part 2, Winter 2022

Below, type the full names of all members of your group below. *By submitting these names, you affirm that you have neither given nor received unauthorized help on this exam.*

```
In [0]: In [0]:
```

Steve Le, Cody Noh, Grace Wu, Josie Rose

```
# TODO
```

Instructions:

This is Part II of the final. It consists of 24 parts organized into three questions. The total point value is 100 points (43% of final exam grade).

Please enter all of your answers either as code or markdown text into the appropriate place a copy of this notebook on CoCalc (similar to homework). Some problems will require Python programming.

Then, you will only submit one version of the midterm, for your entire group, via Gradescope. The Gradescope assignment for the Midterm allows group participation, so make sure you select and enter all members of your group.

Since you are completing this on CoCalc but uploading to Gradescope, we recommend that you wait to upload until you have completed the full exam on CoCalc. At that point, on CoCalc, select File / Download as… / PDF. CoCalc will convert to a PDF that you can download. This entire PDF can then be easily uploaded to Gradescope, and you can manually select where each answer is for each question.

Please be aware, when uploading PDFs and your answers to Gradescope, that page-breaks can accidentally hide text.

Additionally, if you answer questions via code comments (as opposed to markdown text), unless you manually create line breaks in your comment, the full text answer may not be legible. After uploading to Gradescope, please double-check all of your answers to make sure that they are legible and clear.

While screen shots are an acceptable alternative to uploading PDF, due to low resolution, we do not recommend taking photos of your computer screen with your phone, unless absolutely necessary.

Again, we will not grade any material left on CoCalc. In order to receive a grade, you must submit – as a group – to Gradescope. Don't forget to add group members to your submission on Gradescope!

You may use your notes, assignments, slides, readings, solutions, and other resources on our LS 40 BruinLearn site and your CoCalc project (but not elsewhere on the internet).

However, as always, you must show all of your work to receive full credit for each problem.

If you have a clarifying question about the exam at any point during the exam period, post a question on Campuswire "To Instructors and TAs", so our instructional team can help you as quickly as possible. Questions about content or your own progress will not be answered.

For technical glitches with Python, try "Kernel menu > Restart kernel" or Backups in files view first.

In [3]: Out[3]:

In [0]:

In [4]: In [5]:

Gradescope will forbid uploads after 11:30 am Pacific Time on Friday, March 18, 2022. Please plan accordingly.

1. Melts in your mouth, not in your exam

Needing fuel before your exam, you open a single bag of regular M&M candies, pouring the contents out onto your desk. There are 35 candies of the following colors:

Blue: 10

Orange: 9

Green: 1

Yellow: 6

Red: 5

Brown: 4

a. You first wonder whether the probability of getting any given color of M&M is more likely than any other color. What single statistical test you would use to test this hypothesis. If multiple suitable tests are possible, justify your choice. (3 points)

```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats.stats import spearmanr
from scipy.stats.stats import pearsonr
from scipy.stats.stats import linregress
```

```
/tmp/ipykernel_977/2909096273.py:5: DeprecationWarning: Please use `spearmanr` from
the `scipy.stats` namespace, the `scipy.stats.stats` namespace is deprecated.  from
scipy.stats.stats import spearmanr
/tmp/ipykernel_977/2909096273.py:6: DeprecationWarning: Please use `pearsonr` from the
`scipy.stats` namespace, the `scipy.stats.stats` namespace is deprecated.  from
scipy.stats.stats import pearsonr
/tmp/ipykernel_977/2909096273.py:7: DeprecationWarning: Please use `linregress` from
the `scipy.stats` namespace, the `scipy.stats.stats` namespace is deprecated.  from
scipy.stats.stats import linregress
```

```
# TODO
#We would want to use the chi-squared goodness of fit test because the test helps us
compare the observed frequency from our given data to an expected frequency or what
our theoretical frequencies should be. We are trying to determine if our variable (in
this case probability of getting any given color of M&M) comes from a specific
distrubution. This test is also appropriate because our sampling method is random and
our variables are categorical (i.e. the color of candy).
```

b. What would be the observed test statistic for the appropriate test suggested in (a)? [note: in your calculations, do not round to whole M&Ms, fractions of M&Ms are OK] (3 points)

```
# TODO
#Observed test statistic we would want to use is the chi-squared test. Given that each
M&M color has an equal and likely chance of being pulled we can use the chi sqaured
test to calculate our expected frequency for each color
def chi_squared(obs, exp):
 freq = np.sum(((obs-exp)**2)/exp)
 return freq

#To calculate our expected frequency for each color of M&M we would take the total
```


Out[5]: In [6]:

Out[6]:

In [7]: Out[7]:

In [8]: Out[8]:

```
number of M&M's and divide by the total number of colors that we have expected_freq =
35/6
print(expected_freq)
```

```
5.833333333333333
```

```
#For our total chi-squared value we would perform a chi-square test for each color and
sum all of the values together.
print(chi_squared(10, expected_freq))
print(chi_squared(9, expected_freq))
print(chi_squared(1, expected_freq))
print(chi_squared(6, expected_freq))
print(chi_squared(5, expected_freq))
print(chi_squared(4, expected_freq))
```

```
2.9761904761904767
1.7190476190476196
4.004761904761904
0.004761904761904779
0.11904761904761897
0.576190476190476
```

```
chi_obs = chi_squared(10, expected_freq) + chi_squared(9, expected_freq) +
chi_squared(1, expected_freq) +chi_squared(6, expected_freq) +chi_squared(5,
```

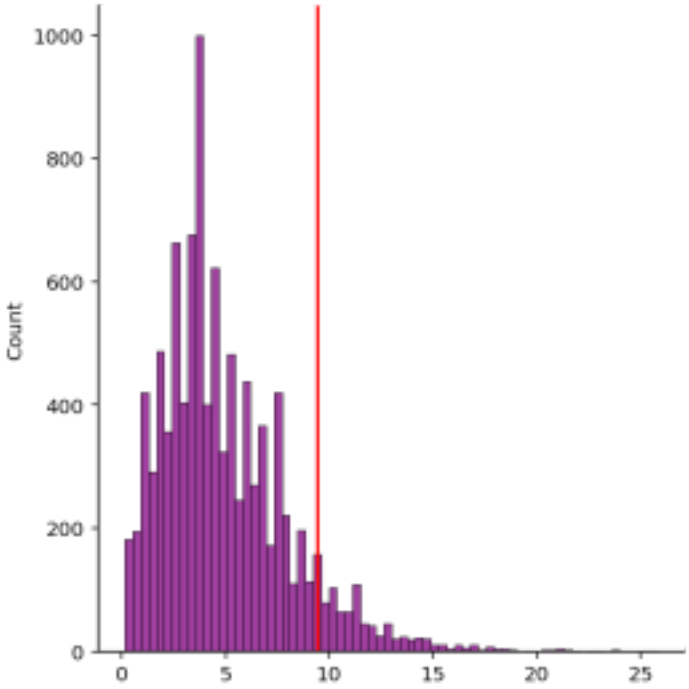
```
expected_freq) + chi_squared(4, expected_freq)
print("The Chi-squared value is", chi_obs)
```

```
The Chi-squared value is 9.399999999999999
```

c. Calculate the probability of observing a test statistic of equal or greater magnitude purely due to random chance under the null hypothesis that frequencies are equal. (3 points)

```
# TODO
results = np.zeros(10000)
null_big_box = 10*["B"]+10*["O"]+10*["G"]+10*["Y"]+10*["R"]+10*["BRW"] for i in range
(10000):
 random = np.random.choice(null_big_box, 35)
 p_blue = np.sum(random == "B")
 p_orange = np.sum(random == "O")
 p_green = np.sum(random == "G")
 p_yellow = np.sum(random == "Y")
 p_red = np.sum(random == "R")
 p_brown = np.sum(random == "BRW")
 chi_val = chi_squared(p_blue, expected_freq)+chi_squared(p_orange,
expected_freq)+chi_squared(p_green, expected_freq) + chi_squared(p_yellow,
expected_freq)+ chi_squared(p_red, expected_freq)+ chi_squared(p_brown, expected_freq)
 results[i]= chi_val

plot = sns.displot(data = results, color = "purple")
plt.axvline(chi_obs, color = "red")
count = np.sum(results >=chi_obs) #Only calculating a one-tail p-value since we are
only trying to find the number of simulations that are equal to or more extreme than
our observed value. Only one direction.Plus chi-square values are only positive so it
would not make sense to calculate a negative one.
p_val = count/10000
print("The p_value for having results greater than or equal to our observed test stat
is", p_val)
```

```
The p_value for having results greater than or equal to our observed test stat is
0.0934
```



In [0]: In [0]:

In [0]:

d. Is the probability calculated in part (c) one-tailed or two-tailed? Why or why not? (3 points)

```
# TODO
#The probability calculated in part (c) will only be one-tailed because our study is trying to determine the number of chi-squared simulations greater than or equal to our observed chi-value. Hence, one direction. Furthermore, chi-square values are only positive becaause we are squaring the values in the calculation so it would not make sense to calculate a two-tailed test because they cannnot be negative. The chi-square value will only be greater than 0 if there is a difference between our observed and expected value as a chi-squared value of 0 tells us that there is no difference.
```

e. What is your inference from the probability calculated in part (c)? (3 points)

```
# TODO
#From our calculated p-value of 0.094, we fail to reject the null hypothesis as it is greater than our significant level of 0.05. Our null hypothesis stated that the probability that our distribution of colors did not differ than our expected distribution or that any difference is due to chance. We can say that our observed distribution of color in the M&M's was due purely to random chance.
```

f. Emotionally invested in this discovery, you show your results to your roommate. Looking at the sorted colors of M&Ms on your desk, she comments, "it seems unlikely that you would draw 10 blue and only 1 green just by chance if the probabilities are truly equal. Can you just calculate that specific probability for me?" Knowing everything you learned in LS40 and based off the preceding analysis, do you comply with her request? If yes, explain how you would calculate the requested probability. If not, defend your choice. (3 points)

```
# TODO
#We can comply with her request by using a NHST and running multiple random simulations in order to find that probility of getting 10 blue and only 1 green. We would use the big-box method as seen above and resample with a sample size of 35. Now instead of calculating the chi-squared value we can count the number of times that the simulations draw 10 blue and only 1 green. Our null hypothesis would state that there is no difference in frequency of our observed and expected color distrubution, but we can manipulate the previous code to now find the number of simulations that resulted in 10 blues with 1 green. From this we could calculate a p-value that would tell us how significantly significant the probabilty of pulling that specific number of colors is.
```

After a night of google searches, you discover that all M&Ms are manufactured at just two different plants in the United States, each packaging a slightly different proportion of colors. Each plant produces approximately 50% of all M&Ms sold in the United States per year. The color frequencies for these two factories are shown in the table below:

```
print("Brown", 35*.205)#Brown Ten
```

```
B+O 8.75
Else 4.375
Tennesee Factory:
Blue 4.585
Orange 4.34
Green 6.930000000000001
Yellow 4.7250000000000005
Red 7.244999999999999
Brown 7.175
```

In [9]: Out[9]:

```
# TODO
#NJ chi square
NJ_chi_obs = chi_squared(10, 8.75) +
chi_squared(9, 8.75) + chi_squared(1, 4.375)+
chi_squared(6, 4.375)+ chi_squared(5, 4.375)+
chi_squared(4, 4.375) print("NJ_chi_obs is",
NJ_chi_obs)
```

In [10]:

```
Ten_chi_obs = chi_squared(10, 4.585) +
chi_squared(9, 4.340) + chi_squared(1, 6.930)
+chi_squared(6, 4.725) + chi_squared(5, 7.245)+
chi_squared(4, 7.175) print("Ten_chi_obs is",
Ten_chi_obs)
```

```
NJ_chi_obs is 3.5142857142857147
Ten_chi_obs is 18.917813832641343
```

Out[10]: In [11]:

h. Calculate the probability of opening a bag of M&Ms and finding the frequency of colors that you observed (or a frequency distribution more extreme), given the expected frequencies from each factory. Your answer should calculate two probabilities, one for each factory. [hint: For your box model, imagine each factory produces 1000 M&Ms of the expected proportions. Refer to homework 7 solutions if you're having trouble coding it.] (6 points)

table.png

But which of these factories sends M&Ms to your local store, you wonder. Determined to find out, you do more statistics.

g. These new factory %s have changed your expectations. Using the same analytic framework as before, calculate two observed test statistics based on the data from your original bag of M&Ms. One test statistic should relate to the expectations from the New Jersey factory and the other test statistic should relate to the expectations from the Tennessee factory. [note: again, do not round to whole M&Ms] (6 points)

```
# TODO
#NJ NHST
NJ_big_box =
250*["B"]+250*["O"]+125*["G"]+125*["Y"]+125*["R"
]+125*["BRW"] results_NJ = np.zeros(10000)
for i in range(10000):
 NJ_random = np.random.choice(NJ_big_box, 35)
 nj_blue = np.sum(NJ_random == "B")
 nj_orange = np.sum(NJ_random == "O")
 nj_green = np.sum(NJ_random == "G")
 nj_yellow = np.sum(NJ_random =="Y")
 nj_red = np.sum(NJ_random =="R")
 nj_brown = np.sum(NJ_random == "BRW")
 NJ_chi_sim = chi_squared(nj_blue, 8.75) +
chi_squared(nj_orange, 8.75) +
chi_squared(nj_green, 4.375) +
chi_squared(nj_yellow, 4.375) +
chi_squared(nj_red, 4.375) +
chi_squared(nj_brown, 4.375)
 results_NJ[i] = NJ_chi_sim

NJ_plot = sns.displot(data = results_NJ, color =
"blue")
plt.axvline(NJ_chi_obs, color = "red")
```

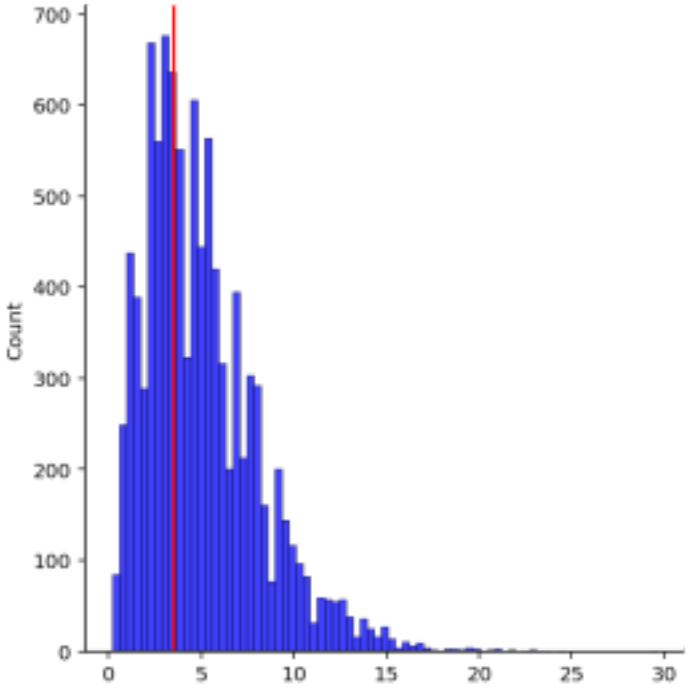
```
print("B+O",35*.25) #Expected frequency for blue
and orange color for NJ print("Else", 35*.125)
#Expected frequency for all the other colors for
NJ print("Tennesee Factory:")
print("Blue", 35*.131)#Blue Ten
print("Orange",35*.124)#Orange Ten
print("Green", 35*.198)#Green Ten
print("Yellow", 35*.135)#Yellow Ten
print("Red", 35*.207)#Red Ten
```

Out[11]:

In [12]: Out[12]:

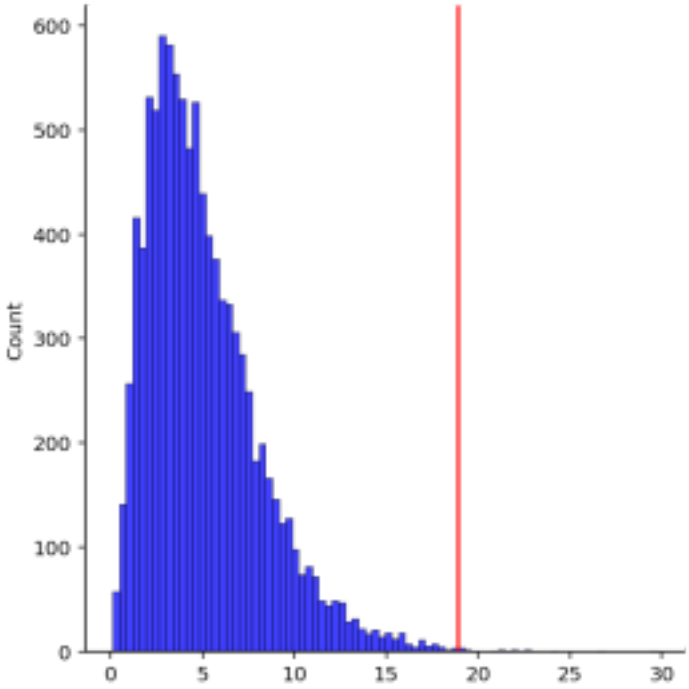
```
NJcount = np.sum(results_NJ >= NJ_chi_obs)
NJ_pval = NJcount/10000
print("The p-value for New Jersey simulation is", NJ_pval)
```

```
The p-value for New Jersey simulation is 0.627
```




```
#Tennessee NHST
Ten_big_box = 131*["B"]+ 124*["O"]+198*["G"]+135*["Y"]+207*["R"]+205*["BRW"]
results_ten = np.zeros(10000)
for i in range(10000):
 Ten_random = np.random.choice(Ten_big_box, 35)
 ten_blue = np.sum(Ten_random == "B")
 ten_orange = np.sum(Ten_random == "O")
 ten_green = np.sum(Ten_random == "G")
 ten_yellow = np.sum(Ten_random == "Y")
 ten_red = np.sum(Ten_random == "R")
 ten_brown = np.sum(Ten_random == "BRW")
 Ten_chi_sim = chi_squared(ten_blue, 4.585) + chi_squared(ten_orange, 4.340) +
chi_squared(ten_green, 6.930) + chi_squared(ten_yellow, 4.725) + chi_squared(ten_red,
7.245) + chi_squared(ten_brown, 7.175)
 results_ten[i] = Ten_chi_sim

Ten_plot = sns.displot(data = results_ten, color = "blue")
plt.axvline(Ten_chi_obs, color = "red")

Tencount = np.sum(results_ten >= Ten_chi_obs)
Ten_pval = Tencount/10000
print("The p-value for Tennessee simulation is", Ten_pval)
```

```
The p-value for Tennessee simulation is 0.0017
```



In [0]: In [0]:

i. Based on your calculations, which factory is more likely to produce a color frequency matching your bag of M&Ms? How certain are you? Can you rule out one of the two factories as the producer? Support your answer only with information from your previous calculations. (3 points)

```
# TODO
#Based on our calcualtions, the New Jersey
factory had a pvalue of 0.6215. This tells us
that we cannot reject the null hypothesis
because it is over the significance level of
0.05. The New Jersey value is much higher than
that of the Tennessee factory which was only
0.0029. This would tell us that there is a
higher chance of getting our bag from the NJ
factory. We can be very certain of this since
the Tennessee factory only prooduced a p-value
of 0.0029 which would tell us that we can
reject the null hypotehsis and say that there
was significant difference between our observed
and expected frequencies. As a result we can
rule out the Tennessee factory at a significant
level, however, there is still a small chance
(about our Tennessee pvalue) that our bag of
M&M's did come from the Tennessee factory. With
random chance, the possibility that either
factory being the producer is possible.
```

j. You present your results to your statistics professor, who proudly claims that "now we know the probability that this bag of M&M's was produced by each factory!" Do you agree with this claim? Why or why not? (3 points)

```
# TODO
#We cannot say anything about the probability
that this bag of M&M's was produced by either
factory. The p-value only tells the likelihood
that a bag of our observed frequency M&M's
would be produced according to the frequencies
of each factory. There is still a chance that
the Tennessee factory could have produced our
M&M bag, but the probability is just lower than
that of the New Jersey factory.
```

2. Nicki Minaj goes to medical school

The following passage was published in the Journal of the American Medical Association (2021, vol. 326, pp. 273– 274) on the effects of two doses of mRNA COVID-19 vaccination on the qualities of sperm of healthy men:

table%201.png

a. On September 13, 2021, musical artist Nicki Minaj tweeted her support of a theory that mRNA vaccines for COVID-19 can decrease male fertility. Her tweet was consistent with the unfounded beliefs of a broad sector of the population, and such concerns were a major source of vaccine hesitancy throughout 2021. The results above, however, suggest that the opposite is true, that mRNA vaccines significantly increase

In [0]:

In [0]: In [0]: In [0]:

In [0]: In [0]:

male fertility (as measured by sperm concentration). Provide two possible interpretations of these results (that is, imagine a U.S. senate panel is asking you what these results mean; this question can be answered without any additional information on the biology involved or the details of the study). (4 points)

```
# TODO
#In the study, the median sperm concentration increased by 4 million/ml with a p value of 0.02. Using a significant threshold level of 0.05, the increase sperm production in the sample can be seen as significant in that there is only a 2% chance that our results were due to pure random chance.

#Similarly the median total motile sperm count also increased by 8 million with a p value of 0.001. The p-value is under the significant threshold value which could then be told to the Senate panel that the findings of the experiment would only be due to pure random chance 0.1% of the time which is a good indication that the varaibles in the experiment could be associated.
```

b. For each of the bolded and underlined numbers in the results section above, classify each as one of the following (not all possible labels will be used): sample mean, sample median, 25th percentile, 75th percentile, effect size, p-value, confidence interval lower bound, confidence interval upper bound, and sample size. (4 points)

```
#Sample Size is 45 men
#25th percentile for baseline sperm concentrationis 19.5 million/ml #75th pecentile of baseline TMSC is 51 million
#Pvalue of TMSC after vaccine is 0.001
```

c. For each p-value identified in part b, give a plausible null hypothesis that is being tested. (3 points)

```
# TODO
#There is no difference in the median TMSC (Total motile sperm count) of the men before and after they recieved their second doses of their respective vaccines.
```

d. For each p-value identified in part b, describe what type of analysis you would use to calculate the p-value if you had the raw data (3 points)

```
# TODO
#To calculate the difference we would use a paired-sample test since we are trying to find the difference in the median of the TMSC for the same sample group, only at a different time. Our null hypothesis of the test would state that the median difference between the two sets of observation is 0 and our study would aim to try to simulate the null hypothesis at a significant level.
```

e. For each p-value identified in part b, identify whether it is likely to be a one-tailed or two-tailed p-value. Explain your reasoning. (3 points)

```
# TODO
#Our p-value in part (b) would most lilely be a one-sided p-value because the table specifcially stated that the p-value was for the increase of TMSC. Including the increase would indicate that a one-sided p-value was used as opposed to a two-sided p-value that would be referenced saying something along the line that there was a "difference" but does not specificy a direction.
```

f. Ideally, the researchers aimed to test whether COVID-19 mRNA vaccines caused a decrease in male fertility (as evidenced by changes in sperm concentration). Provide two important ways that the researchers could have improved their study design to better achieve this goal. (4 points)

```
# TODO
#The researchers can increase their sample size in order to improve their study design. They can also calculate a one-sided p-value in the negative direction that would tell the researchers if a decrease in male fertility (if one existed) would be significant at our standard significance threshold. Paired designed based on matched sampling would also help improve the the study design. They can also improve their study by focusing on just one vaccine rather than using both Pfizer and
```

```
Moderna to reduce any variation that may occur
between the different vaccines.
```

study, and after using Bayes' Theorem, concludes that the probability that mRNA vaccines affect sperm concentration (in any direction) given these data is 0.01%. What would a Bayesian statistician have to account for – that the original researchers did not – which would explain why the two groups arrived at different conclusions? (3 points)

g. Imagine that a Bayesian statistician is given all of the raw data used in this

In [0]:

In [13]: In [14]:

Out[14]: In [16]:

In [17]: Out[17]:

In [42]: Out[42]:

```
# TODO
#Many other things could affect why the Bayesian statistician arrived at a different
conclusion because it uses posterior beliefs to draw conclusions upon their
hypothesis. For example, the diet and lifestyles of the participants would be taken
into account because they could affect the fertility of the men. The Bayesian
```

```
statistician would have to include outside factors that include more than whether the
participants received or didn't recieve the vaccine.
```

3. (S)querulous Correlations

A set of wildlife biologists set out to study the relationship between tail length and tail bushy-ness in UCLA squirrels. Below are sets of x-y measurements (x = length in cm, y = bushy-ness in mm) for each of 11 individuals:

```
X1 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
Y1 = [9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74]
```

a. Calculate the Pearson correlation coefficient for the above set of data. (4 points)

```
# TODO
xarray = np.array(X1)
yarray = np.array(Y1)
print(pearsonr(xarray, yarray))
print("The pearson correlation is", 0.8162365060002428)
```

```
(0.8162365060002428, 0.0021788162369107975)
The pearson correlation is 0.8162365060002428
```

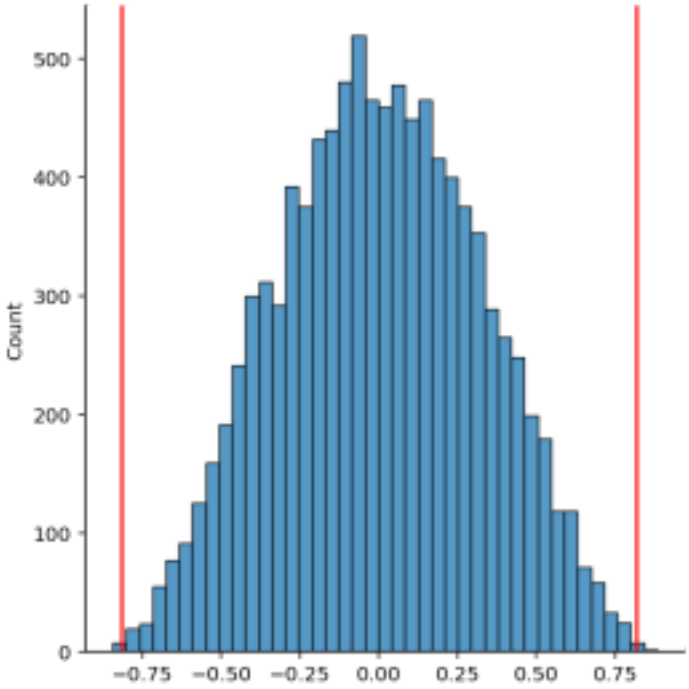
```
df = pd.DataFrame(np.column_stack([X1, Y1]))

df.columns = ["X1", "Y1"]

pobs = pearsonr(df["X1"], df["Y1"])[0]
pobs
```

```
0.8162365060002428
```

b. Use a suitable resampling method to estimate whether there is a significant correlation between tail length and tail bushy-ness. (4 points)

```
# TODO
tail= list(df["X1"])
bushy = list(df["Y1"])
new = np.zeros(10000)
for i in range (10000):
 np.random.shuffle(tail)
 new[i]=pearsonr(tail, bushy)[0]
p=sns.displot(data=new, kde = False)
plt.axvline(pobs, color="red")
plt.axvline(-pobs, color="red")
squirrelcount = sum(new>= pobs) + sum(new<= -pobs)
p_val = squirrelcount/10000
print(p_val)
print("our Pvalue is", p_val)
```

```
0.0007
our Pvalue is 0.0007
```



In [0]:

```
directions, both positive and negative.
One-sided p-value are also rarely used in
calculating this type of test. We are also
trying to test if there is a relationship
between tail-length and tail-bushiness in
either direction, either positive or negative.
Performing a two-sided test would help the
goals of the study by providing evidence for
an association in tail-length and
tail-bushiness.
```

In [64]: Out[64]:

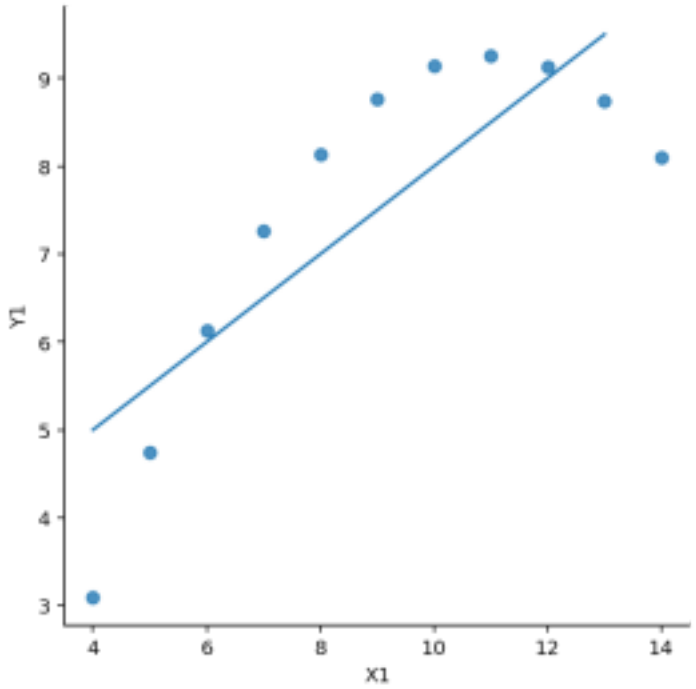
c. In part (b), did you calculate a 1- or 2-sided p-value? Justify your decision in the context of the data, the test, and the research goals. (3 points)

```
# TODO
#We calculated a two-sided p-value because our
statistical significance data goes in both
```

d. Encouraged by the results, the researchers are interested in learning how much bushier UCLA squirrel tails are for every cm of length. Calculate this statistic and provide an appropriate confidence interval. [Assume that tail lengths are measured and reported exactly.] (6 points)

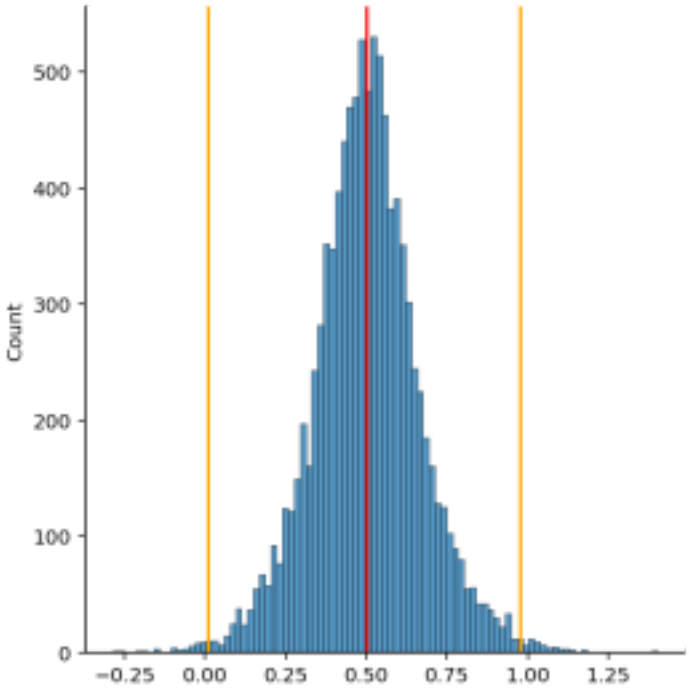
```
# TODO
regress = linregress(df["X1"], df["Y1"])
slope = regress.slope
Y_intercept = regress.intercept
X_plot = np.linspace(4, 13, 11)
Y_plot = slope*X_plot + Y_intercept
plot = sns.lmplot(x = "X1", y = "Y1", data
=df, fit_reg = False) plt.plot(X_plot, Y_plot)
print("Regression slope is", slope)
print("y-intercept is", Y_intercept)
```

```
Regression slope is 0.5000000000000001
y-intercept is 3.000909090909089
```



In [69]: Out[69]:

```
Results=np.zeros(10000)
for i in range(10000):
 Q=df.sample(11, replace=True)
 reg=linregress((Q["X1"]), (Q["Y1"]))
 Results[i]=reg.slope

sns.displot(data=Results)

Results.sort()
Results[49]
Results[9949]
M_upper=(2*.500000000001)-Results[49]
M_lower=(2*0.500000000001)-Results[9949]

plt.axvline(0.500000000001, color="red")
plt.axvline(M_upper, color="orange")
plt.axvline(M_lower, color="orange")


print("The 99% confidence interval is",

(M_lower, M_upper)) The 99% confidence

interval is (0.007659574488085075,

0.9757943925433646)
```



In [0]:

In [28]:

In [29]:

Out[29]: In [36]:

In [38]:

Out[38]: In [45]:

```
#Our regression slope tells us that for every
cm increase in tail length, there is an
increase of an average of about 0.5mm in
tail-bushiness for the UCLA squirrels. Our
calculated 99% confidence interval is (0.007,
0.975). This means that in 99% of our
simulations, this interval will contain the
true value. Since this interval does not
include 0 and is positive we can say that there
```

is a positive correlation between tail size and tail bushiness.

e. A competing set of wildlife biologists from USC decide to replicate the squirrel study and come up with the following measurements from 11 individuals:

```
X2 = [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
Y2 = [7.46, 6.77, 12.74, 7.11, 7.81, 8.84,

6.08, 5.39, 8.15, 6.42, 5.73]
```

With this new set of data,

calculate the correlation coefficient and p-value (as in parts a–b). (8 points)

```
# TODO
xarray2 = np.array(X2)
yarray2 = np.array(Y2)
pearsonr(xarray2, yarray2)
```

```
(0.8162867394895982, 0.002176305279228025)
```

```
df2 = pd.DataFrame(np.column_stack([X2, Y2]))
df2.columns = ["X2", "Y2"]

pobs2 = pearsonr(df2["X2"], df2["Y2"])[0]
pobs2
```

```
0.8162867394895982
```

```
tail2= list(df2["X2"])
bushy2 = list(df2["Y2"])
new2 = np.zeros(10000)
for i in range (10000):
 np.random.shuffle(tail2)
 new2[i]=pearsonr(tail2, bushy2)[0]
p=sns.displot(data=new2, kde = False)
plt.axvline(pobs2, color="red")
plt.axvline(-pobs2, color="red")
squirrelcount2 = sum(new2>= pobs2) + sum(new2<=
-pobs2)
```

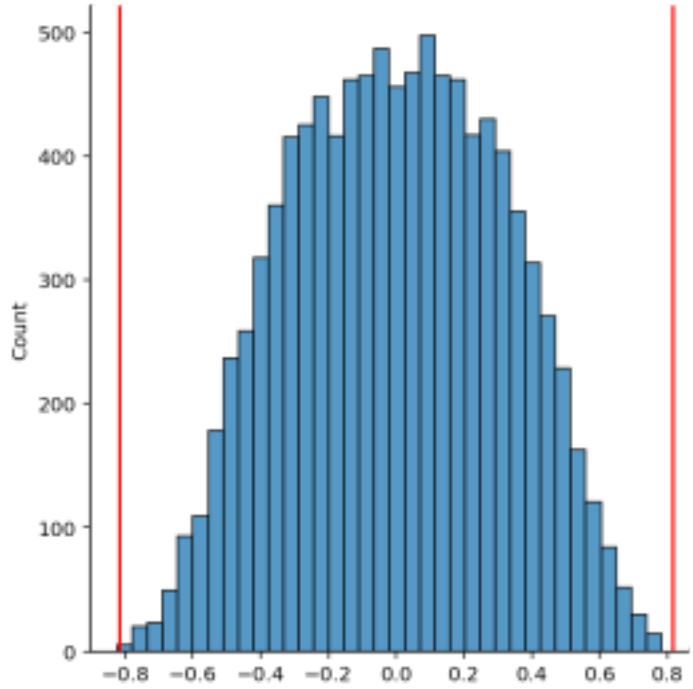


```
p_val2 = squirrelcount2/10000
print(p_val2)
print("our Pvalue is", p_val2)
```

Out[45]:

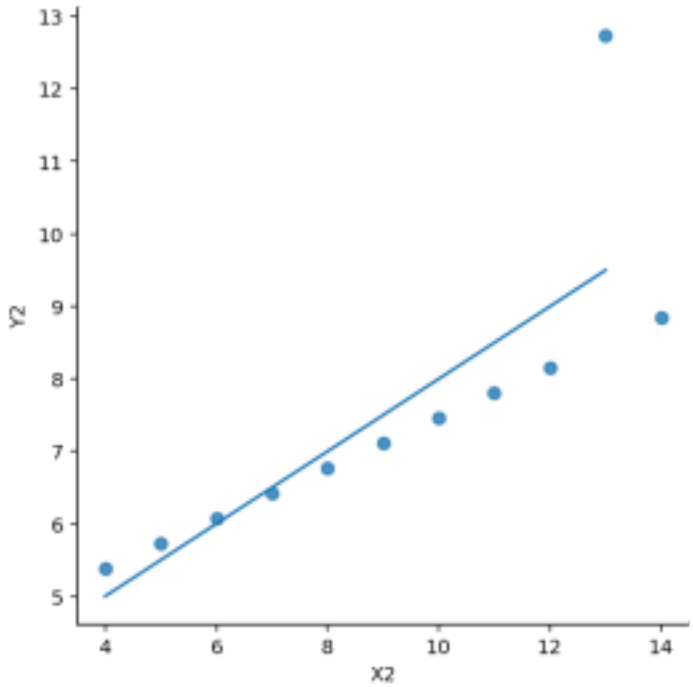
In [41]: Out[41]:

```
0.0001
our Pvalue is 0.0001
```




```
regress2 = linregress(df2["X2"], df2["Y2"])
slope2 = regress2.slope
Y_intercept2 = regress2.intercept
X_plot2 = np.linspace(4, 13, 11)
Y_plot2 = slope2*X_plot2 + Y_intercept2
plot2 = sns.lmplot(x = "X2", y = "Y2", data =df2, fit_reg = False) plt.plot(X_plot2,
Y_plot2)
print("Regression slope is", slope2)
print("y-intercept is", Y_intercept2)
```

```
Regression slope is 0.4997272727272729
y-intercept is 3.002454545454544
```



In [0]:

```
squirrls which fit the line of regression
better. Since from the graph of the UCLA
squirrels was was not even linearly related
based on the visualization of the data, using
pearson's correlation and linear regression in
the first place would not be appropriate. Since
the data is obviously non-linear, calculating
linear regression would give us results that
are not representative of the data. Therefore,
it would not make sense to compare the results
of the UCLA squirrels to the U$C squirrels
based off of the correlation coefficients.
```

```
In [0]: In [0]:
```

f. Sharing their results, the UCLA and USC scientists come together and conclude that the relationship between tail length and tail bushy ness in squirrels is the same on both campuses. You don't agree. Convince them otherwise, using at least one appropriate graph to support your position. (10 points)

```
# TODO
#The relationsip between tail length and tail
bushiness in the squirrels on both campuses are
not the same because the the data points for
the UCLA squirrels are not closely related to
the line of regression compared the U$C
```

g. Thanks to your persuasive evidence, the two teams of researchers conclude that the relationship between tail length and bushy-ness is not the same between the two schools. In order for them to not make the same mistake again, what final lesson would you impart? (5 points)

```
# TODO
#Before you go into any sort of analysis of
data, make sure you are looking at the
distrubition of the data to make sure that the
appriate statisitcal test is being used. In
addition, you cannot assume that there is a
relationship between any variable based on any
data since even correlation does not equal
causation. Especially with a small sample size
of 11 squirrels, using just a significant
p-value does not even indicate a significance
relationship. The researchers can strengthen
their experiment by calculating a power value
as well as increasing sample size or performing
more tests. LESSON: Alwyas visualize the data
before you do anything.
```

Congratulations, you are done! When you are ready to submit, please choose File > Save and Download As > PDF and then upload to Gradescope.