HW-06 SOLUTIONS

QUESTION 1

Part D) Are X and Y independent or dependent? Fully justify your answer in the cell below using LaTeX and the mathematical definition of independence.

SOLUTION:

They are dependent. Can show this with any counterexample. For example, notice ($p_{X,Y}$(1,2) = 1/12) ≠ (p_X(1)p_Y(2) = 5/12 . 5/12)

QUESTION 2:

Part A) If Cov(X, Y) = 0, what does this tell us about the random variables X and Y?

Solution: They don't have a linear relationship. It does NOT tell us

whether or not they are independent.

Part B) Given the following joint pmf for discrete random variables X and Y :

	Y = 0	Y = 1	Y = 2
X = 0	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{8}$
Y = 1	$\frac{1}{8}$	$\frac{1}{6}$	$\frac{1}{6}$

- i). Calculate Cov(X, Y).
- ii). Calculate P(X, Y)

Show all steps for both parts using Markdown and LaTeX in the cell below:

SOLUTION to part (i):

$Cov(X,Y) = E\,[\,(X-\mu X)\,(Y-\mu Y)\,] = E[XY] - E[X]E[Y]$

To find $\mu X = E[X]$ and $\mu Y = E[Y]$ it will be easiest to start by finding the marginal distributions of X and Y:

x	0	1
$pX(x)$	13/24	11/24

$\Rightarrow E[X] = 0(13/24) + 1(11/24) = 11/24$

y	0	1	2
$pY(y)$	7/24	5/12	7/24

$\Rightarrow E[Y] = 0(7/24) + 1(5/24) + 2(7/24) = 1$

Option 1:
Using $E[(X-\mu X)\,(Y-\mu Y)] = \sum x \sum y (x-\mu X)\,(y-\mu Y)\,pX,Y(x,y)$

=
(0−11/24)(0−1)1/6+(0−11/24)(1−1)1/4+(0−11/24)(2−1)1/8+(1−11/24)(0−1)1/8+(1−11/24)(1−1)1/6+(1−11/24)(2−1)/16

=124

Option 2:

Using $E[XY] - E[X]E[Y]$

$E[XY]$**=(0)(0)1/6+(0)(1)1/4+(0)(2)1/8+(1)(0)1/8+(1)(1)1/6+(1)(2)1/6 = 1/2**

$\Rightarrow Cov(X,Y) = E[XY] - E[X]E[Y] = 1/2 - 11/24(1) = 1/24$

SOLUTION to part (ii):

$\varrho(X,Y) = Cov(X,Y) \,/\, \sigma X \sigma Y$

$\sigma 2X = E[X2] - (\mu X)2$

$E[X2] = 11/24$

$\Rightarrow \sigma 2X = 11/24 - (11/24)\text{^}2 = 143/576$

$\Rightarrow \sigma X = 143/576$_____$\surd$

Similarly:

$E[Y2]$ **= 1(5/12)+2^2(7/24) = 19/12**

$\Rightarrow \sigma 2_Y$ **= 19/12 − (1)2 = 7/12**

$\Rightarrow \sigma_Y$ **= 7/12**$\_\_\_\_\sqrt{}$

$\Rightarrow \varrho(X,Y) = Cov(X,Y) / \sigma X \sigma Y$ **= 1/24(143/576**$\sqrt{}$**)(7/12**$\sqrt{}$**) = 2/31001**$\_\_\_\_\_\sqrt{}$ **≈0.11**

Part C) This part is **NOT** related to the parts above.
Suppose you're only given the following information about two joint random variables X and Y :

μX = 6, μY = 5, σ2 X = 4, σ2Y = 9 and E[XY] = 27

For each of the quantities below, calculate if you have enough information, showing all steps. If not, explain what additional info you'd need.

i). Cov(X, Y)

ii). Cov(Y , X)

iii). P(X, Y)

Answer all parts in the ONE markdown cell below, fully justifying your answer:

Solution

a).
$Cov(X,Y)$ **=** $E[XY]$ **−** $E[X]E[Y]$ **= 27−(6)(5) = −3**

b).
$Cov(Y,X)$ **=** $E[YX]$ **−** $E[Y]E[X]$ **= 27−(6)(5) = −3**

c).
$\varrho(X,Y)$**=**$Cov(X,Y)$ **/** $\sigma X \sigma Y$ **= −3 / (2)(3) = −1/2**

0.1 (2 pts) Problem 4

If we're trying to predict the results of the Clinton vs. Trump 2016 presidential race:

i). What is the population of interest?

ii). What is the sampling frame?

Give both of your answers in the same below in Markdown.

SOLUTION to 4i: The population is all people who will vote/ voted in the 2016 US presidential election.

SOLUTION to 4 ii: Anyone who has a phone (they need to be reached by random digit dialing). In addition, survey respondents are excluded if they are deemed unlikely to vote or not elligible for voting in the upcoming election.

PROBLEM 5

Part D i). Make a **frequency** histogram of simulations. This is a histogram of the sampling distribution of Trump's proportion advantage in Pennsylvania.

Hint: You should use the plt.hist function in your code.

Make sure to include a title as well as axis labels. You can do this using plt.title, plt.xlabel, and plt.ylabel.

ii). Based on your simulation, what is the probability that a random sample of 1500 will correctly predict that Trump wins Pennsylvania? (i.e. what proportion of these simulations predict a Trump victory?) Assign your answer to prob_penn_1500_random_correct

In [ ]:
```
# Part (i):
...
    plt.hist(simulations) ;
    plt.title('Pennsylvania');
    plt.ylabel('# of Simulations');
    plt.xlabel('Sampling Distribution Advantage');
# your code for the histogram above here. The code below plots a red marker at the mean:
plt.scatter(simulations.mean(), -1, marker='^', color='red', s=500)
```

In [ ]:
```
# Part (ii):
```

```
prob_penn_1500_random_correct =   sum(simulations > 0)/len(simulations) # SOLUTION

prob_penn_1500_random_correct
```

PROBLEM 6

Part B Create a plot of **overlaid DENSITY** histograms of the following: - The new sampling distribution of Trump's proportion advantage in Pennsylvania using these biased samples - The sampling distribution of the unbiased samples from Problem 5D (plotted as a density, not a frequency histogram)

Include 2 markers (of different colors) with the sample means for each distribution (see 5D for code how to do this). The colors of the markers should correspond to the colors of the density histograms.

Make sure to give your plot a title, label the x and y axes and include a legend. Use the parameter alpha to adjust the transparency of each histogram.

In [ ]: ...

```
plt.hist(biased_simulations, density=True) ;
plt.hist(simulations, density=True, alpha=0.5) ;
plt.title('Biased Sampling of Pennsylvania') ;
plt.ylabel('# of Simulations');
plt.xlabel('Sampling Distribution Advantage');

plt.scatter(biased_simulations.mean(), -1, marker='^', color='blue', s=500)
plt.scatter(simulations.mean(), -1, marker='^', color='orange', s=500)
```

e) Summarize the findings from these simulations:

i). Based on your simulations, what was the **chance of error** in correctly predicting that Trump wins using the **unbiased** samples of 1500 people from each state? Many people, even well educated ones, assume that this number should be 0%. After all, how could a non-biased sample be wrong? Give a mathematical explanation as to why it isn't 0% (or close to 0%). This is the type of incredibly important intuition we hope to develop in you throughout this class and your future data science coursework.

ii). What was the chance of error in predicting the results using the **biased** samples and how different is it from your answer in part(i)? Recall, we only biased the samples by 0.5%. However, even a bias this small in the percentages can lead to a much larger chance of error in prediction of the final result.

SOLUTION

Numbers will vary based on answers above.
i) Should be approximately 1-.69 = .31

ii). Should be approximately 1-.46 = 0.54
Main idea is that there is inherent variation in a sample, so even if we used an unbiased sampling approach, we won't be correct 100% of the time.

PROBLEM 7

Part B Compare your observations from 7a to your observations in 6d. Did the chance of error increase or decrease in each case and why? What do these changes imply about the impact of sample size on the sampling error and on the bias?

SOLUTION

The larger sample size has decreased the spread in the Trump lead histogram. The unbiased sample of 5000 now correctly predicts Trump a winner over 80% of the time (i.e. the chance of error is close to 20%)

The shift in the histogram remains the same size. Since the sampling error has decreased, the change in the error of predicting Trump being a winner has increased from about 54% to about 56%.

Part C Is it possible to correctly predict Trump's victory with less than 1% error using **unbiased sampling?** Rerun the simulation (in each of the 4 states) with increasing sample sizes and 100,000 simulations to determine if you can find an approximate minimum sample size (it doesn't have to be exact) such that the probability of correctly predicting Trump's victory is at least 99% (assuming your sample is unbiased).

In [ ]: ...

```
np.mean([trump_wins(10000) for i in range(100000)])
np.mean([trump_wins(20000) for i in range(100000)])
np.mean([trump_wins(30000) for i in range(100000)])
print("31000 is ", np.mean([trump_wins(31000) for i in range(100000)]))

# your code above this line.
 # output the number of samples you used to get to at least 99% accuracy.
```

31000 is 0.99002

Part D Is it possible to correctly predict Trump's victory with less than 1% error using **biased sampling?** Use the code cell below to rerun the simulation (in each of the 4 states) with increasing sample sizes. What happens to the probability of error? Explain in the markdown cell below.

SOLUTION

We have seen that the predictions are driven by the sample size and the size of the bias:

No, it is not possible to get to that level of accuracy with biased sampling.
Unfortunately, if there is bias, then the predictions are close to the biased estimate. If the bias pushes the prediction from one candidate (Trump) to another (Clinton), then we will have an incorrect biased estimate of the outcome (which we have in this case).