Raul G. Martinez (PID: A12461871)

DSE 210: Probability and statistics Winter 2020

Take-Home Final Part 1 — Sampling

1. *(2 points) A box contains 9 red marbles and 1 blue marbles. Nine hundred random draws are made from this box, with replacement. What is distribution of the number of red marbles seen, roughly?*

The probability that a red marble is seen from this box in each draw is p = 0.9

We can use the binomial distribution for n = 900, but first lets define the expected value and variance:

Expected Value: μ = np = (900)(0.9) = 810

Variance: σ^2 = np(1-p) = (900)(0.9)(1-0.9) = 81

Therefore, the distribution of the number of red marbles can be defined as ***N*****(810, 81)**; where the standard deviation is $\sigma = 9$

2. *(5 points, 1 each) A dartboard is partitioned into 20 wedges of equal size, numbered 1 through 20. Half the wedges are painted red, and the other half are painted black. Suppose 100 darts are thrown at the board, and land at uniformly random locations on it.*

(a) Let X_i be the number of darts that fall in wedge i. What are $E(X_i)$ and $var(X_i)$?

Use the binomial distribution for n = 100 to find the number of darts X_i, where the probability of a dart to fall in wedge *i* can be defined as p = 1/20. Hence:

$$E(X_i) = np = (100)(1/20) = \mathbf{5}$$

$$var(X_i) = \sigma^2_{Xi} = np(1-p) = (100)(1/20)(1-1/20) = \mathbf{19/4}$$

(b) Using a normal approximation, give an upper bound on X_i that holds with 95% confidence.

From part (a) we know the standard deviation by simply taking the square root of the variance. Also, the critical value Z^* for 95% confidence interval is 1.96, then using the formula below we can calculate the upper bound for the normal approximation as:

$$\text{95\% Confidence Interval} = E(X_i) \pm Z^* \sigma_{Xi} = 5 \pm (1.96)\left(\sqrt{\frac{19}{4}}\right) = \mathbf{5 \pm 4.3}$$

Rounding the number of darts for the confidence interval to whole numbers from 4.3 to 4, the upper bound that holds 95% confidence is where $\mathbf{X_i \leq 9}$

Let Z_r be the number of darts that fall on red wedges, let Z_b be the number of darts that fall on black wedges, and let $Z = |Z_r - Z_b|$ be the absolute value of their difference. We would like to get a 99% confidence interval for Z. To do this, define

$$Y_i = \begin{cases} 1 & \text{if } i\text{th dart falls in red region} \\ -1 & \text{if } i\text{th dart falls in black region} \end{cases}$$

and notice that $Z_r - Z_b$ can be written as $Y_1 + Y_2 + \cdots + Y_{100}$, the sum of independent random variables.

(c) What are $E(Y_i)$ and $var(Y_i)$?

Considering that the probability of y = 1 and y = -1 for Y_i is the same at 1/2, and using the equation below we can find the expectation and variance for Y_i as:

$$E(Y_i) = \sum_y y \Pr(Y_i = y)$$

$$E(Y_i) = (1)(1/2) + (-1)(1/2) = \mathbf{0}$$

$$\text{var}(Y_i) = E(Y^2_i) - (E(Y_i))^2 = (1)^2 (1/2) + (-1)^2 (1/2) - (0)^2 = 1/2 + 1/2 = \mathbf{1}$$

(d) Using the central limit theorem, we can assert that $Z_r - Z_b$ is approximately a normal distribution. What are the parameters of this distribution?

The parameters for the distribution for Y_i are $E(Y_i) = 0$ and $\text{var}(Y_i) = 1$, then for n = 100 we can use the central limit theorem to define $Sn = Z_r - Z_b$ where the expectation and variance parameters are:

$E(Sn) = nE(Y_i) = (100)(0) = 0$; $\text{var}(Sn) = n\text{var}(Y_i) = (100)(1) = 100$

Therefore, the approximation for the normal distribution for Sn looks like $\boldsymbol{N}(\mathbf{0, 100})$

(e) Give a 99% confidence interval for Z.

The critical parameter Z^* for 99% confidence is 2.576, we also have the standard deviation for Yi from the calculated variance above and we can use the expected value for the number of darts X_i to find the confidence interval for Z as:

$$\text{99\% Confidence Interval} = E(Y_i) \pm Z^* \sigma_{Yi} = 0 \pm (2.576)\left(\sqrt{100}\right) = \mathbf{0 \pm 25.76}$$

Since negative numbers are not practical here we can exclude everything below zero, also we can round the upper bound to whole numbers within the interval from 25.76 to 25. **Hence, we get [0, 25] for the 99% confidence interval.**

3. *(2 points, 1 each) You have hired a polling agency to determine what fraction of San Diegans like sushi. Unknown to the agency, the actual fraction is exactly 0.5.*

   The agency is going to poll a random subset of the population and return the observed fraction of sushi-lovers. How far off would you expect their estimate to be (i.e. what standard deviation) if:

   a) *they poll 100 people?*

   We can use the binomial distribution for n = 100 to find the standard deviation, but first lets define the expected value and variance for a fraction:

   Expected Value: μ = p = 0.5

   Variance: σ^2 = p(1-p)/n = (0.5)(1-0.5)/100 = 0.0025

   $$\textbf{Standard Deviation: } \boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}^2} = \sqrt{\mathbf{0.0025}} = \mathbf{0.05}$$

   b) *they poll 2500 people?*

   Again, we can use the binomial distribution for n = 2500 to find the standard deviation, but first lets define the expected value and variance for a fraction:

   Expected Value: μ = p = 0.5

   Variance: σ^2 = p(1-p)/n = (0.5)(1-0.5)/2500 = 0.0001

   $$\textbf{Standard Deviation: } \boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma}^2} = \sqrt{\mathbf{0.0001}} = \mathbf{0.01}$$

4. *(3 points) In a certain city, there are 100,000 people age 18 to 24. A random sample of 500 of these people is drawn, of whom 194 turn out to be currently enrolled in college. Estimate the percentage of all persons age 18 to 24 in the city who are enrolled in college. Give a 95.5% confidence interval for your estimate.*

   The observed fraction, out of n = 500, is:

   $\hat{p}$ = 194/500 = 0.388

   Now we can approximate the expected value, variance, and standard deviation as:

   Expected Value: μ = p = 0.388

   Variance: σ^2 = p(1-p)/n = (0.388)(1-0.388)/500 = 0.00047

   Standard Deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{0.00047}$ = 0.021

   Referencing the Z table for 95.5% confidence interval the critical value Z^* equals 2.0, therefore we can find the confidence interval as:

   95.5% Confidence Interval = $\mu \pm Z^* \sigma$ = 0.388 ± (2.0)(0.022) = 0.388 ± 0.044 = **38.8% ± 4.4%**

5. *(2 points) A survey research company uses random sampling to estimate the fraction of residents of Austin, Texas, who watch Spanish-language television. They are satisfied with the estimate they get using a sample size of 1,000 people.*

   They then want to also estimate this fraction for Dallas, which has similar demographics to Austin, but twice the population. What sample size would be suitable for Dallas?

   Since both Austin and Dallas have similar demographics, the fraction of residents p who watch Spanish-language television is similar. What matters is the sample size, not the overall population size. Therefore, **the sample size suitable for Dallas would be similar to Austin which is 1,000 people.**

6. *(2 points, 1 each) A box contains many pieces of papers with numbers on them. 100 random draws are made from the box, with replacement, and the sum of the draws is 297.*

   (a) Can you estimate the average of the numbers in the box?

   Using the sum and the number of draws, the average can be estimated as: $\boldsymbol{\mu}$ **= 297/100 = 2.97**

   (b) Can you give a confidence interval for your estimate, based on the information so far?

   Considering the variance of a sum we can estimate the variance and standard deviation like:

$$\sigma^2 = n = 100;\ \sigma = \sqrt{n} = \sqrt{100} = 10$$

   Referencing the Z table for a 95% confidence interval the critical value Z^* equals 1.96, therefore we can find the confidence interval as:

$$95.5\% \text{ Confidence Interval} = \mu \pm Z^* \sigma = 2.97 \pm (1.96)(10) = \mathbf{2.97 \pm 19.6}$$

DSE 210: Probability and statistics Winter 2020

Take-Home Final Part 2 — Hypothesis testing

7. *(3 points, 1 each) The Public Health Service studied the effects of smoking on health, in a large sample of representative households. For men and women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.*

(a) Was this a controlled experiment or an observational study?

Observational Study – The representative households were sampled and they assigned themselves to the study groups based on their habits.

(b) Why did they study men and women and the different age groups separately?

There are two facts (provided in our class notes) for smoking observational studies that answer this question:

For Sex: Men are more likely to smoke than women, and are more likely to get heart disease.

For Age: Older people have different smoking habits and are more at risk for the disease.

(c) The conclusion seems to be that you shouldn't start smoking, but once you've started, you shouldn't stop. Comment.

Correct, I agree that the conclusion above reflects the statement described on this problem. However, I believe the conclusion seems to be very general and it can be more persuasive if it provides more information about the subjects sampled; for example, if the statements above can be compared between men and women across different ages (i.e. 21-30, 31-40, 50+). Since this is not a controlled experiment is hard to draw a definite conclusion, but the more data you gather as an observer the more correlations will be available to find a handful of them, that when combined with some context (i.e. sex and age), might explain the data better.

8. *(3 points, 1 each) In 10,000 tossings, a coin came up heads 5,400 times. Should we conclude that the coin is biased?*

(a) Formulate the null hypothesis and alternative hypothesis.

Null Hypothesis, H_0: The data comes from a fair coin with probability p = 0.5

Alternative Hypothesis, H_1: The data comes from a biased coin *(b) Compute the z statistic and the p-value.*

First find the mean and standard deviation for the number of tossings n = 10000 and probability p = 0.5 for a fair coin:

Mean (expected) = np = (10000)(0.5) = 5000

Stddev = $\sqrt{\mathrm{np}(1-\mathrm{p})} = \sqrt{(10000)(0.5)(1-0.5)} = 50$

Now we can compute the z statistic and p-value, where the observed value in this experiment is 5400. Then:

z = $\frac{\text{observed} - \text{expected}}{\text{standard deviation}} = \frac{5400 - 5000}{50} = \mathbf{8}$

therefore, our **p-value is < 0.00001**

(c) What do you conclude?

Since our p-value is very small (< 0.00001) there is extremely strong evidence against the null hypothesis. Hence, we can reject the null hypothesis H_0 and accept the alternative hypothesis H_1, which basically tells that **our data comes from a biased coin.**

9. *(4 points) A die is rolled 100 times. The total number of spots is 368 instead of the expected 350. Can this be explained as chance variation, or is the die loaded?*

Null Hypothesis, H_0: The data comes from a fair die where each side has p = 1/6

Alternative Hypothesis, H_1: The data comes from a loaded die

The probability of each spot is p = 1/6 and n = 100. Let's find the standard deviation, z statistic, and p-value to answer this question:

Stddev = $\sqrt{\mathrm{np}(1-\mathrm{p})} = \sqrt{(100)(1/6)(1-1/6)} = 3.73$

z = $\frac{\text{observed} - \text{expected}}{\text{standard deviation}} = \frac{368 - 350}{3.73} = 4.83$

Our p-value is < 0.00001, therefore there is very strong evidence against the null hypothesis and we can conclude that **the die is loaded.**

10. *(2 points) Other things being equal, which is better for the null hypothesis: a higher p-value or a lower p-value?*

As stated in our class notes: "The p-value is the probability of seeing a value (at least) this extreme under the null. A small p-value is evidence against the null"

The idea behind hypothesis testing is that we want to design the experiment in a way that we are trying to discredit the null hypothesis H_0, therefore is better for us that the null hypothesis has a low p-value. A high p-value for H_0 means that we don't have enough evidence reject H_0 and to accept our alternative hypothesis H_1, which is the one we normally want to likely be true.

11. *(4 points, 2 each) The National Household Survey on Drug Abuse was conducted in 1985 and 1992. In each year, a simple random sample of 700 people was used.*

    (a) Among persons age 18 to 25, the percentage of marijuana users dropped from 21.9% to 11.0%. Is this difference real, or a chance variation?

    X1 = sample of marijuana users in 1985, with mean = (700)(0.219) = 153.3

    X2 = sample of marijuana users in 1992, with mean = (700)(0.110) = 77.0

    Null Hypothesis, H_0: The means of the two distributions (surveys in 1985 and in 1992) are the same

    Calculate the standard deviation under the null hypothesis:

    X1 is roughly normally distributed, with standard deviation ~ $\sqrt{700}$

    X2 is roughly normally distributed, with standard deviation ~ $\sqrt{700}$

    Under the null hypothesis, X1 – X2 is normally distributed with mean zero and standard deviation

$$\sqrt{700+700} \sim 37.4$$

    Now, calculate the z statistic for X1 – X2:

$$z = \frac{\text{observed} - \text{expected}}{\text{standard deviation}} = \frac{153.3 - 77.0}{37.4} = \mathbf{2.04}$$

    The difference is real, the probability of observing this difference under the null hypothesis is about 2%. In other words, our p-value is ~0.0205 for a significant level of 0.05 in a one-tailed hypothesis test. This is strong evidence against the null hypothesis, hence we can reject it.

(b) Among persons age 18 to 25, the percentage of cigarette smokers dropped from 36.9% to 31.9%. Is this difference real, or a chance variation?

X1 = sample of cigarette smokers in 1985, with mean = (700)(0.369) = 258.3

X2 = sample of cigarette smokers in 1992, with mean = (700)(0.319) = 223.3

Null Hypothesis, H_0: The means of the two distributions (surveys in 1985 and in 1992) are the same

Calculate the standard deviation under the null hypothesis:

X1 is roughly normally distributed, with standard deviation ~ $\sqrt{700}$

X2 is roughly normally distributed, with standard deviation ~ $\sqrt{700}$

Under the null hypothesis, X1 – X2 is normally distributed with mean zero and standard deviation

$$\sqrt{700 + 700} \sim 37.4$$

Now, calculate the z statistic for X1 – X2:

$$z = \frac{\text{observed} - \text{expected}}{\text{standard deviation}} = \frac{258.3 - 223.3}{37.4} = \mathbf{0.94}$$

The difference is only a chance variation, the probability of observing this difference under the null hypothesis is about 17%. In other words, our p-value is ~0.1736 for a significant level of 0.05 in a one-tailed hypothesis test. This is not enough evidence against the null hypothesis, hence we cannot reject it.

12. *(4 points) A random sample of 1000 freshmen at public universities were asked how many hours they worked each week (for pay). The average number of hours turned out to be 12.2, with a standard deviation of 10.5. A similar survey at private universities had an average of 9.2 hours, with a standard deviation of 9.9. Is the difference between these two averages due to chance?*

X1 = avg. number of working hours in public universities, this is equal to 12.2

X2 = avg. number of workings hours in private universities, this is equal to 9.2

Null Hypothesis, H_0: The means of the two distributions (work hours in public and private universities) are the same

Calculate the standard deviation under the null hypothesis:

X1 is roughly normally distributed, with standard deviation $\sigma_1 = 10.5/\sqrt{1000} = 0.33$

X2 is roughly normally distributed, with standard deviation $\sigma_2 = 9.9/\sqrt{1000} = 0.31$

Under the null hypothesis, X1 – X2 is normally distributed with mean zero and standard deviation:

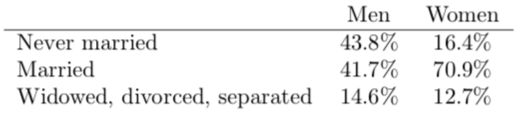
$$\sigma = \sqrt{\sigma_1{}^2 + \sigma_2{}^2} = \sqrt{(0.33)^2 + (0.31)^2} = 0.45$$

Now, calculate the z statistic for X1 – X2:

$$z = \frac{12.2 - 9.2}{0.45} = 6.66$$

The difference is not due to chance, is real. The likelihood of observing this difference under the null hypothesis is more than 6 standard deviations from the mean. In other words, our p-value is < 0.00001 for a significant level of 0.05 in a one-tailed hypothesis test. This is strong evidence against the null hypothesis, hence we can reject it.

13. *(4 points) A survey was conducted to determine the distribution of marital status by sex for persons age 25-29 in Wyoming. A random sample of 103 people was chosen, of whom 48 were men and 55 were women. The following results were obtained:*



	Men	Women
Never married	43.8%	16.4%
Married	41.7%	70.9%
Widowed, divorced, separated	14.6%	12.7%

Are the distributions really different for men and women?

Null Hypothesis, H₀: The distribution for martial status by sex of persons age 25-29 in Wyoming is the same

Define a table summarizing the observed and expected values:

	Observed		Total	Expected	
	Men	Women		Men	Women
Never Married	21	9	30 (29.1%)	14	16
Married	20	39	59 (57.3%)	27.5	31.5
Widowed, divorced, separated	7	7	14 (13.6%)	6.5	7.5
Total	48	55	103	48	55

Now we can compute the χ^2 statistic for the data:

$$\chi^2 = \sum_{outcomes} \frac{(observed\ frequency - expected\ frequency)^2}{expected\ frequency}$$

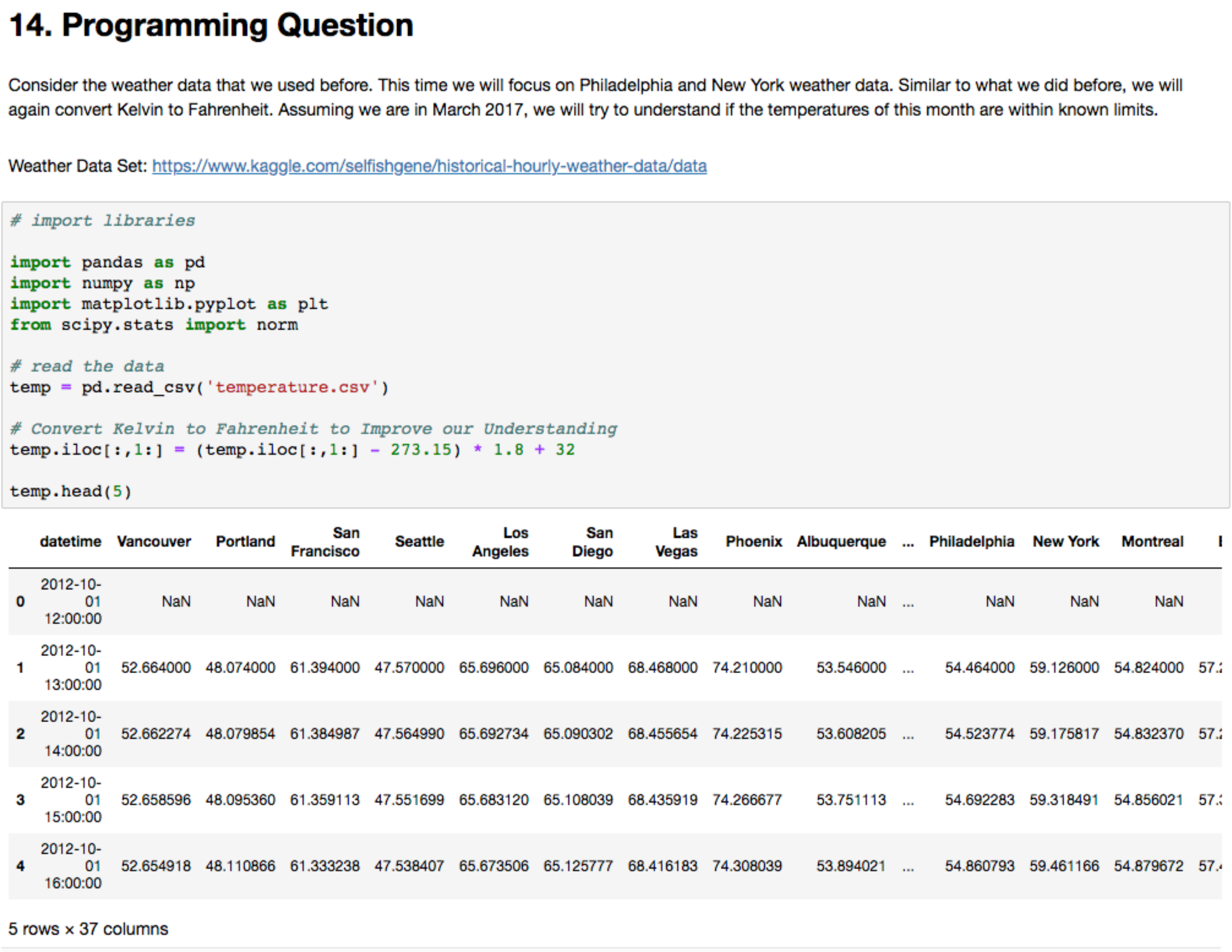
$$\chi^2 = \frac{(21-14)^2}{14} + \frac{(20-27.5)^2}{27.5} + \frac{(7-6.5)^2}{6.5} + \frac{(9-16)^2}{16} + \frac{(39-31.5)^2}{31.5} + \frac{(7-7.5)^2}{7.5} = \mathbf{10.5}$$

And the degrees of freedom (df) for the two-way table, with 3 rows and 2 columns, can be found as:

(r-1)(c-1) = (3-1)(2-1) = 2

Using the table for χ^2 critical values, the distribution for df = 2 and a critical value of ~10.5 has a probability **~ 0.5%** under the null hypothesis. Therefore, we can reject the null hypothesis and conclude that **the distributions are indeed different for men and women.**

14. ***Programming Question:*** *(10 points, 2.5 points each)*
Consider the weather data that we used before. This time we will focus on Philadelphia and New York weather data. Similar to what we did before, we will again convert Kelvin to Fahrenheit. Assuming we are in March 2017, we will try to understand if the temperatures of this month are within known limits.



14. Programming Question

Consider the weather data that we used before. This time we will focus on Philadelphia and New York weather data. Similar to what we did before, we will again convert Kelvin to Fahrenheit. Assuming we are in March 2017, we will try to understand if the temperatures of this month are within known limits.

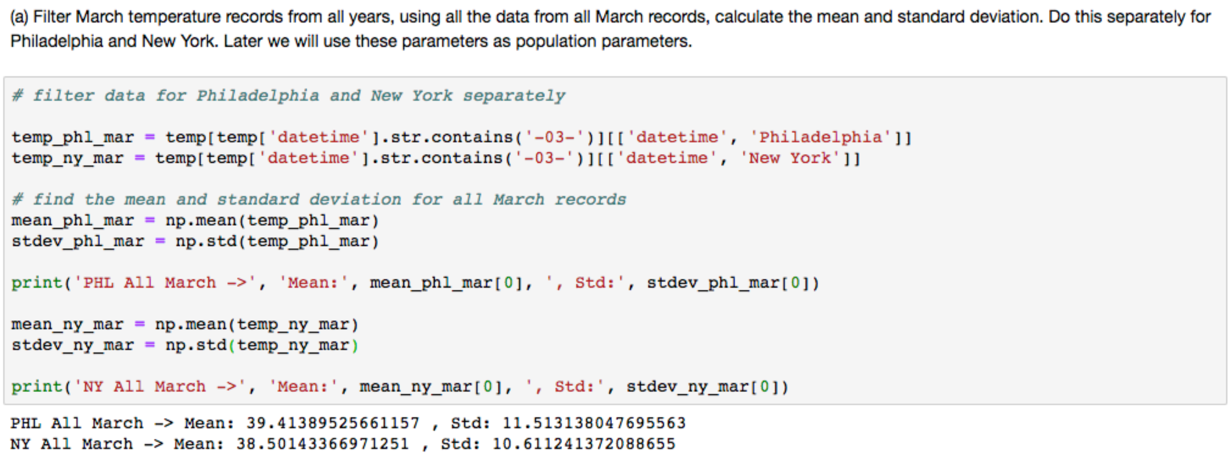
Weather Data Set: https://www.kaggle.com/selfishgene/historical-hourly-weather-data/data

```
# import libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# read the data
temp = pd.read_csv('temperature.csv')

# Convert Kelvin to Fahrenheit to Improve our Understanding
temp.iloc[:,1:] = (temp.iloc[:,1:] - 273.15) * 1.8 + 32

temp.head(5)
```

	datetime	Vancouver	Portland	San Francisco	Seattle	Los Angeles	San Diego	Las Vegas	Phoenix	Albuquerque	...	Philadelphia	New York	Montreal	I
0	2012-10-01 12:00:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	
1	2012-10-01 13:00:00	52.664000	48.074000	61.394000	47.570000	65.696000	65.084000	68.468000	74.210000	53.546000	...	54.464000	59.126000	54.824000	57.2
2	2012-10-01 14:00:00	52.662274	48.079854	61.384987	47.564990	65.692734	65.090302	68.455654	74.225315	53.608205	...	54.523774	59.175817	54.832370	57.2
3	2012-10-01 15:00:00	52.658596	48.095360	61.359113	47.551699	65.683120	65.108039	68.435919	74.266677	53.751113	...	54.692283	59.318491	54.856021	57.3
4	2012-10-01 16:00:00	52.654918	48.110866	61.333238	47.538407	65.673506	65.125777	68.416183	74.308039	53.894021	...	54.860793	59.461166	54.879672	57.4

5 rows × 37 columns

(a) Filter March temperature records from all years, using all the data from all March records, calculate the mean and standard deviation. Do this separately for Philadelphia and New York. Later we will use these parameters as population parameters.

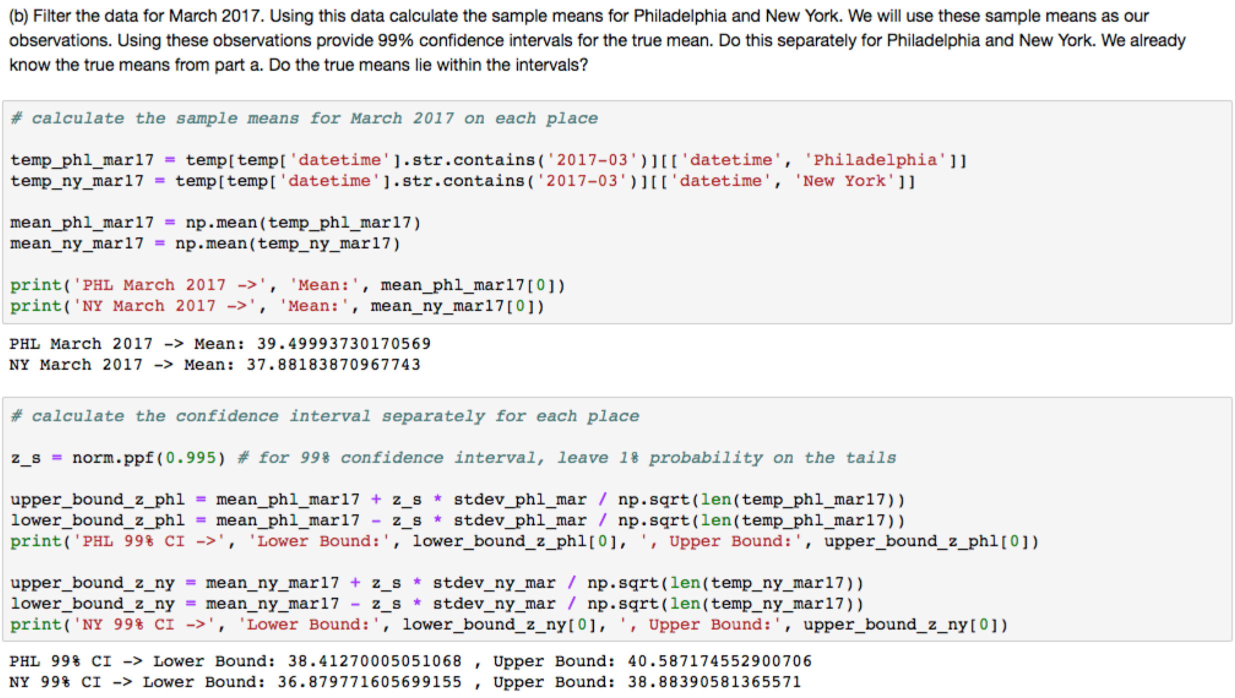


(a) Filter March temperature records from all years, using all the data from all March records, calculate the mean and standard deviation. Do this separately for Philadelphia and New York. Later we will use these parameters as population parameters.

```
# filter data for Philadelphia and New York separately

temp_phl_mar = temp[temp['datetime'].str.contains('-03-')][['datetime', 'Philadelphia']]
temp_ny_mar = temp[temp['datetime'].str.contains('-03-')][['datetime', 'New York']]

# find the mean and standard deviation for all March records
mean_phl_mar = np.mean(temp_phl_mar)
stdev_phl_mar = np.std(temp_phl_mar)

print('PHL All March ->', 'Mean:', mean_phl_mar[0], ', Std:', stdev_phl_mar[0])

mean_ny_mar = np.mean(temp_ny_mar)
stdev_ny_mar = np.std(temp_ny_mar)

print('NY All March ->', 'Mean:', mean_ny_mar[0], ', Std:', stdev_ny_mar[0])
```

```
PHL All March -> Mean: 39.41389525661157 , Std: 11.513138047695563
NY All March -> Mean: 38.50143366971251 , Std: 10.611241372088655
```

(b) Filter the data for March 2017. Using this data calculate the sample means for Philadelphia and New York. We will use these sample means as our observations. Using these observations provide 99% confidence intervals for the true mean. Do this separately for Philadelphia and New York. We already know the true means from part a. Do the true means lie within the intervals?



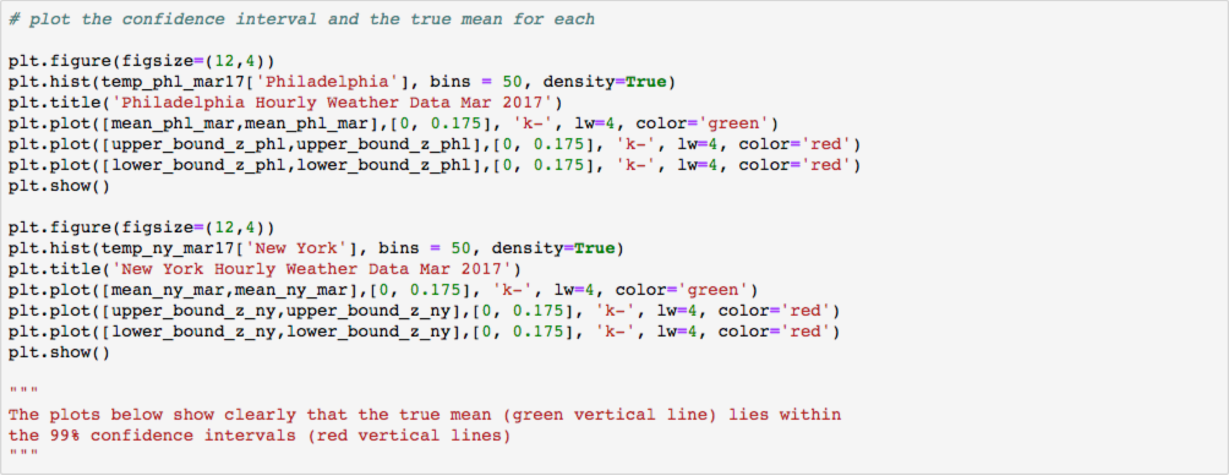
(b) Filter the data for March 2017. Using this data calculate the sample means for Philadelphia and New York. We will use these sample means as our observations. Using these observations provide 99% confidence intervals for the true mean. Do this separately for Philadelphia and New York. We already know the true means from part a. Do the true means lie within the intervals?

```
# calculate the sample means for March 2017 on each place

temp_phl_mar17 = temp[temp['datetime'].str.contains('2017-03')][['datetime', 'Philadelphia']]
temp_ny_mar17 = temp[temp['datetime'].str.contains('2017-03')][['datetime', 'New York']]

mean_phl_mar17 = np.mean(temp_phl_mar17)
mean_ny_mar17 = np.mean(temp_ny_mar17)

print('PHL March 2017 ->', 'Mean:', mean_phl_mar17[0])
print('NY March 2017 ->', 'Mean:', mean_ny_mar17[0])
```

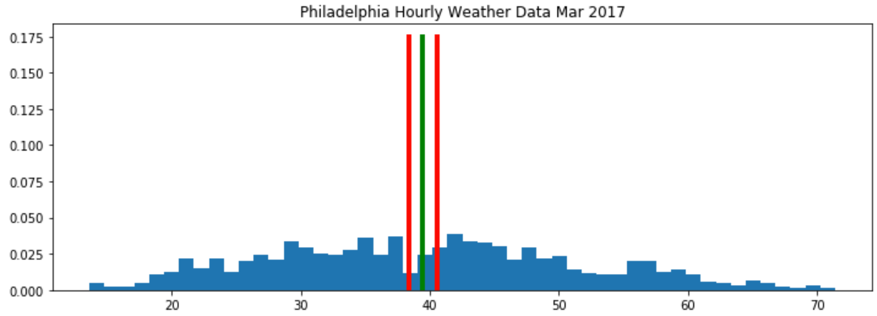
```
PHL March 2017 -> Mean: 39.49993730170569
NY March 2017 -> Mean: 37.88183870967743
```

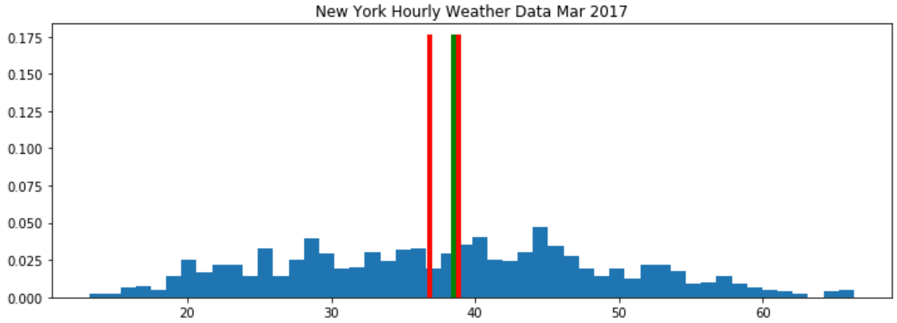
```
# calculate the confidence interval separately for each place

z_s = norm.ppf(0.995) # for 99% confidence interval, leave 1% probability on the tails

upper_bound_z_phl = mean_phl_mar17 + z_s * stdev_phl_mar / np.sqrt(len(temp_phl_mar17))
lower_bound_z_phl = mean_phl_mar17 - z_s * stdev_phl_mar / np.sqrt(len(temp_phl_mar17))
print('PHL 99% CI ->', 'Lower Bound:', lower_bound_z_phl[0], ', Upper Bound:', upper_bound_z_phl[0])

upper_bound_z_ny = mean_ny_mar17 + z_s * stdev_ny_mar / np.sqrt(len(temp_ny_mar17))
lower_bound_z_ny = mean_ny_mar17 - z_s * stdev_ny_mar / np.sqrt(len(temp_ny_mar17))
print('NY 99% CI ->', 'Lower Bound:', lower_bound_z_ny[0], ', Upper Bound:', upper_bound_z_ny[0])
```

```
PHL 99% CI -> Lower Bound: 38.41270005051068 , Upper Bound: 40.587174552900706
NY 99% CI -> Lower Bound: 36.879771605699155 , Upper Bound: 38.88390581365571
```



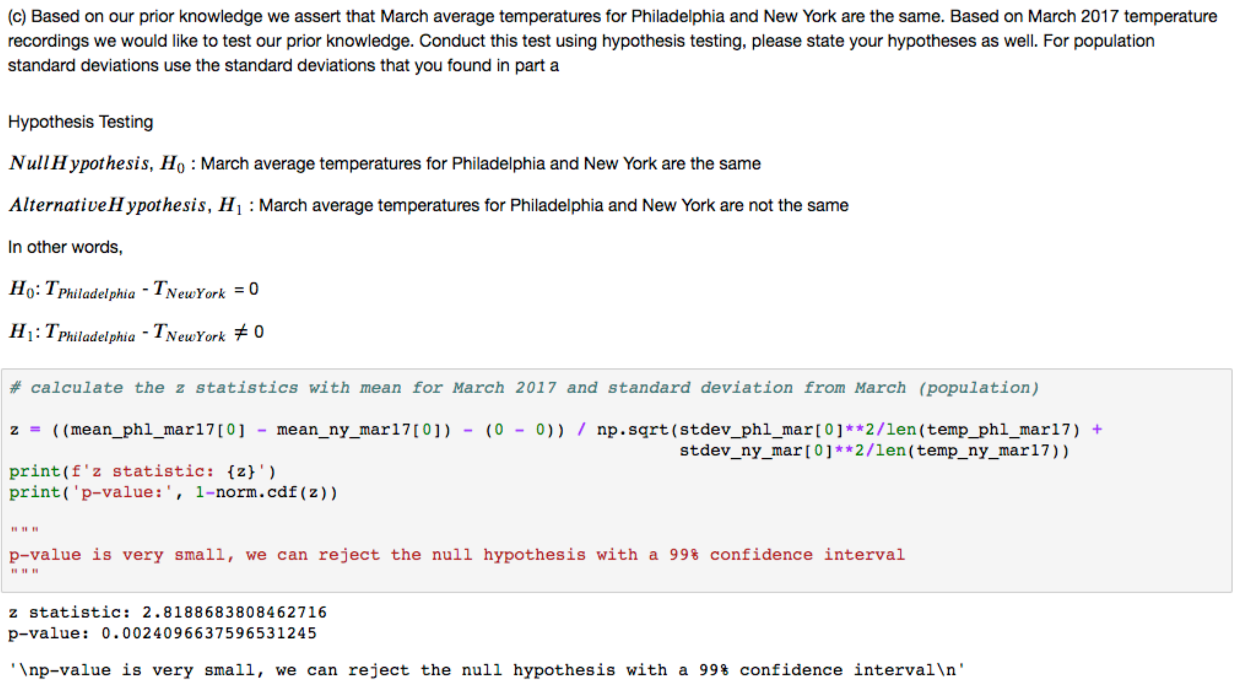
```
# plot the confidence interval and the true mean for each

plt.figure(figsize=(12,4))
plt.hist(temp_phl_mar17['Philadelphia'], bins = 50, density=True)
plt.title('Philadelphia Hourly Weather Data Mar 2017')
plt.plot([mean_phl_mar,mean_phl_mar],[0, 0.175], 'k-', lw=4, color='green')
plt.plot([upper_bound_z_phl,upper_bound_z_phl],[0, 0.175], 'k-', lw=4, color='red')
plt.plot([lower_bound_z_phl,lower_bound_z_phl],[0, 0.175], 'k-', lw=4, color='red')
plt.show()

plt.figure(figsize=(12,4))
plt.hist(temp_ny_mar17['New York'], bins = 50, density=True)
plt.title('New York Hourly Weather Data Mar 2017')
plt.plot([mean_ny_mar,mean_ny_mar],[0, 0.175], 'k-', lw=4, color='green')
plt.plot([upper_bound_z_ny,upper_bound_z_ny],[0, 0.175], 'k-', lw=4, color='red')
plt.plot([lower_bound_z_ny,lower_bound_z_ny],[0, 0.175], 'k-', lw=4, color='red')
plt.show()

"""
The plots below show clearly that the true mean (green vertical line) lies within
the 99% confidence intervals (red vertical lines)
"""
```





```
'\nThe plots below show clearly that the true mean (green vertical line) lies within \nthe 99% confidence intervals
(red vertical lines)\n'
```

(c) Based on our prior knowledge we assert that March average temperatures for Philadelphia and New York are the same. Based on March 2017 temperature recordings we would like to test our prior knowledge. Conduct this test using hypothesis testing, please state your hypotheses as well. For population standard deviations use the standard deviations that you found in part a.



(c) Based on our prior knowledge we assert that March average temperatures for Philadelphia and New York are the same. Based on March 2017 temperature recordings we would like to test our prior knowledge. Conduct this test using hypothesis testing, please state your hypotheses as well. For population standard deviations use the standard deviations that you found in part a

Hypothesis Testing

$NullHypothesis, H_0$: March average temperatures for Philadelphia and New York are the same

$AlternativeHypothesis, H_1$: March average temperatures for Philadelphia and New York are not the same

In other words,

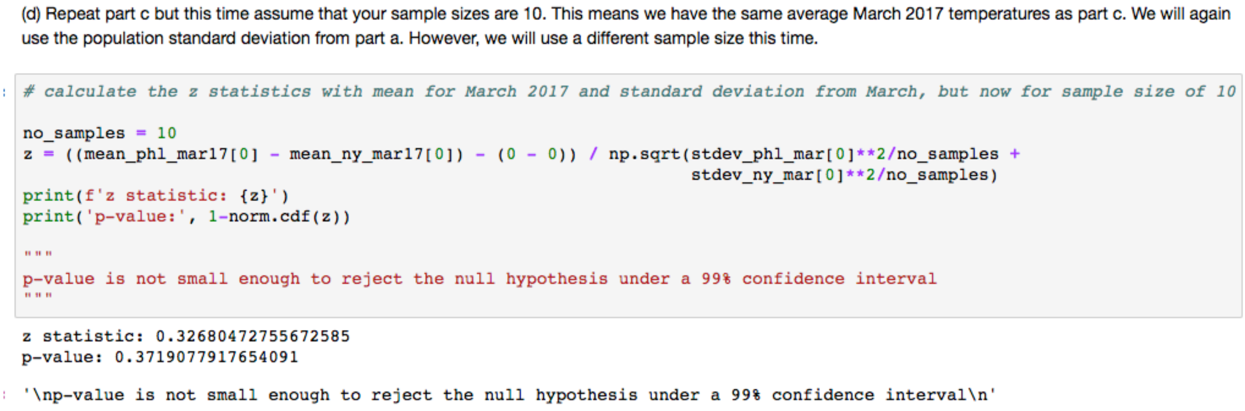
$H_0: T_{Philadelphia}$ - $T_{NewYork}$ = 0

$H_1: T_{Philadelphia}$ - $T_{NewYork}$ $\neq$ 0

```
# calculate the z statistics with mean for March 2017 and standard deviation from March (population)

z = ((mean_phl_mar17[0] - mean_ny_mar17[0]) - (0 - 0)) / np.sqrt(stdev_phl_mar[0]**2/len(temp_phl_mar17) +
                                                                 stdev_ny_mar[0]**2/len(temp_ny_mar17))
print(f'z statistic: {z}')
print('p-value:', 1-norm.cdf(z))

"""
p-value is very small, we can reject the null hypothesis with a 99% confidence interval
"""
```

```
z statistic: 2.8188683808462716
p-value: 0.0024096637596531245

'\np-value is very small, we can reject the null hypothesis with a 99% confidence interval\n'
```

(d) Repeat part c but this time assume that your sample sizes are 10. This means we have the same average March 2017 temperatures as part c. We will again use the population standard deviation from part a. However, we will use a different sample size this time.



(d) Repeat part c but this time assume that your sample sizes are 10. This means we have the same average March 2017 temperatures as part c. We will again use the population standard deviation from part a. However, we will use a different sample size this time.

```
# calculate the z statistics with mean for March 2017 and standard deviation from March, but now for sample size of 10

no_samples = 10
z = ((mean_phl_mar17[0] - mean_ny_mar17[0]) - (0 - 0)) / np.sqrt(stdev_phl_mar[0]**2/no_samples +
                                                                 stdev_ny_mar[0]**2/no_samples)
print(f'z statistic: {z}')
print('p-value:', 1-norm.cdf(z))

"""
p-value is not small enough to reject the null hypothesis under a 99% confidence interval
"""
```

```
z statistic: 0.32680472755672585
p-value: 0.3719077917654091

'\np-value is not small enough to reject the null hypothesis under a 99% confidence interval\n'
```

DSE 210: Probability and statistics Winter 2020

Take-Home Final Part 3 — General Questions

15. *(6 points, 3 each) A smart-phone consists of 10 major components, each independently faulty with probability 1 − p. If any component is faulty, the phone is damaged. Five phones are manufactured independently in sequence. Find the probability that:*

    Write your answers as powers of p and 1 − p.

    (a) the last phone manufactured is the first damaged,

    We can use the geometric distribution to calculate the probability, where the parameters are: the last phone manufactured n = 5 and the probability of a phone being damaged considering all components is p = 10(1-p); then:

$$\Pr(X = n) = p(1-p)^{n-1}$$

$$\Pr(X = 5) = 10(1-p)\big(1-10(1-p)\big)^{5-1} \; = \mathbf{10(1-\boldsymbol{p})(10\boldsymbol{p}-9)^4}$$

    (b) exactly two are damaged.

    Use the binomial distribution where the parameters are n = 5, k = 2, and p = 10(1-p) to estimate the probability as:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\Pr(X = 2) = \frac{5!}{2!\,(5-2)!}\big(10(1-p)\big)^2\big(1-10(1-p)\big)^{5-2} = (10)(100)(1-p)^2(10p-9)^3$$
$$= \mathbf{1000(1-\boldsymbol{p})^2(10\boldsymbol{p}-9)^3}$$

16. *(6 points, 2 each) A fair coin with P(heads) = 0.5 and a biased coin with P(heads) = 0.75 are placed in an urn. One of the two coins is picked at random and tossed twice. Find the probability:*

(a) of observing two heads,

The probability of picking either coin is 0.5, use conditional probability to solve this as:

Pr(two H) = Pr(two H | fair coin) Pr(fair coin) + Pr(two H | biased coin) Pr(biased coin)

= (0.5)(0.5)(0.5) + (0.75)(0.75)(0.5) = 0.406 -> **40.6%**

(b) that the biased coin was picked if two heads are observed.

We can use Bayes rule to solve this question:

$$\text{Pr (biased coin} \mid \text{two H)} = \frac{\text{Pr (two H} \mid \text{biased coin) Pr (biased coin)}}{\text{Pr(two H)}}$$

$$= \frac{(0.75)(0.75)(0.5)}{0.406} = 0.693 \rightarrow \mathbf{69.3}\%$$

(c) Qualitatively explain your answer to part (b) in 1-2 sentences.

The probability of picking a biased coin given that two heads were observed (posterior) is 69.3%. Bayes theorem allowed us to define this probability by considering prior knowledge and conditions related to the event, such as the probability of observing two heads given that is a biased coin (likelihood), the probability of picking a biased coin (prior), and the probability of observing two heads (evidence).

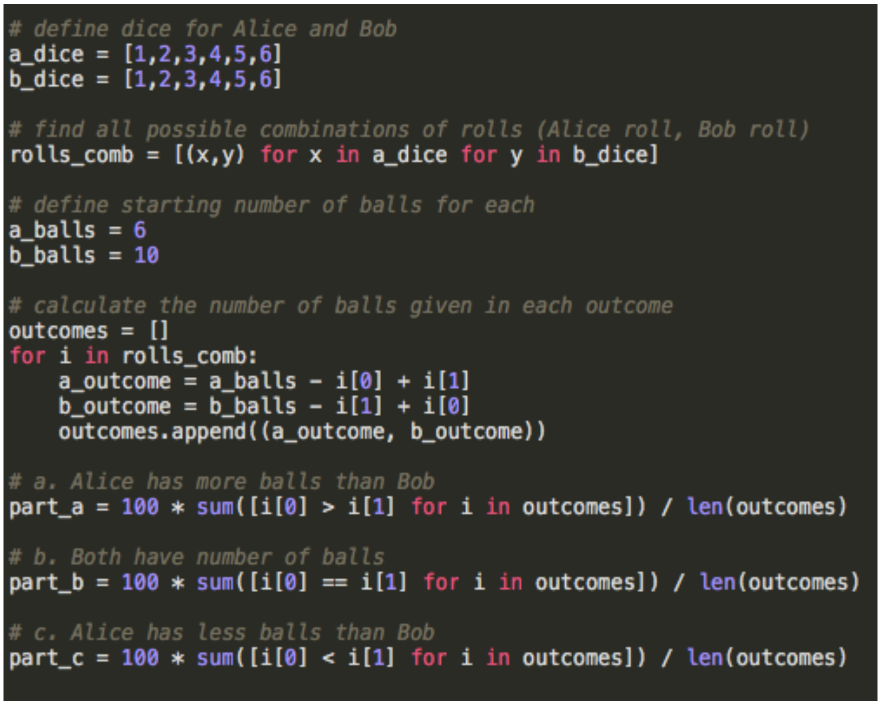
17. *(6 points, 2 each) Alice has 6 balls and Bob has 10. Each of them rolls an independent fair die and gives the other as many balls as their roll's outcome. For example, if Alice rolls 2 and Bob rolls 5, they will end up with 6-2+5=9 and 10-5+2=7 balls respectively. Find the probability that Alice ends up with:*

    (a) strictly more balls than Bob, **16.67%**

    (b) the same number of balls as Bob, **11.11%**

    (b) strictly fewer balls than Bob. **72.22%**

Python code used to solve the problem:



```
# define dice for Alice and Bob
a_dice = [1,2,3,4,5,6]
b_dice = [1,2,3,4,5,6]

# find all possible combinations of rolls (Alice roll, Bob roll)
rolls_comb = [(x,y) for x in a_dice for y in b_dice]

# define starting number of balls for each
a_balls = 6
b_balls = 10

# calculate the number of balls given in each outcome
outcomes = []
for i in rolls_comb:
    a_outcome = a_balls - i[0] + i[1]
    b_outcome = b_balls - i[1] + i[0]
    outcomes.append((a_outcome, b_outcome))

# a. Alice has more balls than Bob
part_a = 100 * sum([i[0] > i[1] for i in outcomes]) / len(outcomes)

# b. Both have number of balls
part_b = 100 * sum([i[0] == i[1] for i in outcomes]) / len(outcomes)

# c. Alice has less balls than Bob
part_c = 100 * sum([i[0] < i[1] for i in outcomes]) / len(outcomes)
```

18. *(6 points, 2 each) n balls are tossed at random into n bins, so that each ball is equally likely to fall in any bin, and different balls are independent of each other. For example, for n = 3, balls 1, 2, and 3, may fall in bins 2, 2, and 1, respectively. Express each of the following in terms of n.*

(a) The probability that bin 1 is empty.

The probability that any ball falls into any of the bins is 1/n, and hence the probability that any ball falls into bin 1 is the same. Then if we take the complement as 1-1/n this now becomes the probability that bin 1 is empty. Considering there are n balls, we now have to take the product for every ball. Therefore, the probability that bin 1 is empty can be expressed as:

$$\Pr(bin\ 1\ is\ empty) = \left(1-\frac{1}{n}\right)^{n}$$

(b) The expected value E(X), where X is the number of empty bins, e.g., for the n = 3 illustration above, just bin 3 is empty, hence X = 1.

Considering the expected value for a binomial distribution, we can use the probability calculated in part (a) and the number of n bins to find the expected value for the random variable X as follows:

$$p = \Pr(bin\ 1\ is\ empty) = \left(1-\frac{1}{n}\right)^{n}$$

$$\boldsymbol{E(X)\ =\ np =\ n\left(1-\frac{1}{n}\right)^{n}}$$

(c) The probability that bins 1 and 2 are empty.

We can use the probability of a binomial distribution where the parameters are p (probability from part (a) that a single bin is empty), n = n (number of bins), and k = 2 (number of chosen bins); need to find all possible combinations for which we can pick bins 1 and 2 from the total.

$$\Pr(X=k) = \binom{n}{k} p^{k}(1-p)^{n-k}$$

$$\Pr(X=2) = \frac{n!}{2!\,(n-2)!}p^{2}(1-p)^{n-2}\ = \boldsymbol{\frac{n(n-1)}{2}\left(1-\frac{1}{n}\right)^{2n}\left(1-\left(1-\frac{1}{n}\right)^{n}\right)^{n-2}}$$

19. *(5 points) Let X and Y be independent random variables with expectations 1 and 2, and variances 3 and 4, respectively. Find the variance of XY .*

Using the properties for independence we can find the variance of the product for X and Y as:

$$Var(XY)\ = [E(X)]^{2}\ Var(Y) + [E(Y)]^{2}\ Var(X) + Var(X)\ Var(Y)$$

$$Var(XY)\ = (1)^{2}\ (4) + (2)^{2}\ (3) + (3)\ (4)\ = 4 + 12 + 12 = \mathbf{28}$$

20. *(3 points, 1 each) Which distribution would you prefer for the following tasks? and why?*

(a) The number of times you would win against Magnus Carlsen in chess when played n games, given that probability of you winning is p in each game.

Binomial Distribution: Because every game played is an independent replicate that yields a binomial outcome, the processed is being repeated a specified number of times, and the probability of success (winning in this case) is the same for every repetition.

We can define a random variable X, where R = Raul and M = Magnus as:

$$\mathrm{X} = \begin{cases} 1 & \text{R wins} \\ 0 & \text{M wins} \end{cases}$$

Therefore, when played n games with probability of Raul winning p. The expected number of times Raul wins becomes:

$$E(X=1) = np$$

(b) Number of shoots Kobe Bryant had to make until seeing his first goal. (Let p be the probability of securing a goal)

Geometric Distribution: Because every shot Kobe Bryant does (RIP) can be considered a Bernoulli Trial (either he makes a goal or not), every trial is independent of each other, and with the geometric distribution we can model the probability of the trials until the first success (goal) is observed.

Here, the expected number of shots he has to make until seeing his first goal (with probability p) is:

$$E(X) = 1/p$$

(c) To determine the probability of exactly 4 storms occurring in 2021 at Toronto, given that on an average 2 storms occur each year.

Poisson Distribution: Because the probability of each storm is a discrete event that occurs randomly in a given interval of time (per year in this case). Also, if we define the number of storms as a random variable we can see that there is theoretically no finite upper limit as it can take any positive integer value.

For example, define the random variable X = number of storms and Poisson parameter λ = storms per year. Hence, the probability can be defined as:

$$\Pr(\mathrm{X} = 4) = \mathrm{e}^{-\lambda}\frac{\lambda^4}{4!}\text{, where } \lambda = 2$$

21. *(5 points) Random variable X is distributed Poisson, and P (X = 2) = P (X = 4). Find P (X = 3).*

Use the Poisson distribution formula to find P(X = 2) and P(X = 4):

$\Pr(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$, therefore:

$\Pr(X = 2) = e^{-\lambda}\frac{\lambda^2}{2!}$, $\Pr(X = 4) = e^{-\lambda}\frac{\lambda^4}{4!}$

We can equal them to each other to find the "number of events per interval" parameter λ

$$e^{-\lambda}\frac{\lambda^2}{2!} = e^{-\lambda}\frac{\lambda^4}{4!} \rightarrow \lambda^2 = 12 \rightarrow \lambda = \sqrt{12}$$

Now find P(X = 3):

$$\Pr(X = 3) = e^{-\sqrt{12}}\frac{\sqrt{12}^3}{3!} = 0.2169 \rightarrow \mathbf{21.69}\%$$

22. *(6 points, 2 each) A computer manufacturer produces 2000 chips, each with independent defect probability 0.001. Using the Poisson approximation for the number X of defective chip, find:*

(a) The Poisson parameter for X

The Poisson parameter (λ) is the expected number of defective chips, therefore if there are 2000 chips where each has probability 0.001, and the expected value is:

$$\lambda = (2000)\,(0.001) = \mathbf{2}$$

(b) P(X > 1)

Using the Poisson formula $\Pr(\mathrm{X} = \mathrm{x}) = \mathrm{e}^{-\lambda}\frac{\lambda^{\mathrm{x}}}{\mathrm{x}!}$, where $\lambda = 2$:

$$\Pr(X > 1) = 1 - \Pr(X = 1) - \Pr(X = 0)$$

$$\Pr(X > 1) = 1 - \Pr(X = 1) - \Pr(X = 0) = 1 - \mathrm{e}^{-2}\frac{2^1}{1!} - \mathrm{e}^{-2}\frac{2^0}{0!} = 1 - 2\mathrm{e}^{-2} - \mathrm{e}^{-2} = \mathbf{1 - 3e^{-2}}$$

(c) P(X ≤ 3)

Using the Poisson formula $\Pr(\mathrm{X} = \mathrm{x}) = \mathrm{e}^{-\lambda}\frac{\lambda^{\mathrm{x}}}{\mathrm{x}!}$, where $\lambda = 2$:

$$\Pr(X \leq 3) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) + \Pr(X = 3)$$

$$\Pr(X \leq 3) = \mathrm{e}^{-2}\frac{2^0}{0!} + \mathrm{e}^{-2}\frac{2^1}{1!} + \mathrm{e}^{-2}\frac{2^2}{2!} + \mathrm{e}^{-2}\frac{2^3}{3!} = e^{-2}\left(1 + 2 + 2 + \frac{8}{6}\right) = \mathbf{\frac{19}{3}}\boldsymbol{e^{-2}}$$

23. *(4 points) The quadratic function $f : IR^3 \rightarrow IR$ given by*

$$f(x) = 4x^2{}_1 + x1x2 + 9x2^2 - x2x3 + 16x^2{}_3$$

can be written in the form of x^T M x for some symmetric matrix M . What is M ?

Define x and x^T as:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},\ x^T = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

Then, we can re-write the quadratic function as follows:

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 4{x_1}^2 + x_1x_2 + 9{x_2}^2 - x_2x_3 + 16{x_3}^2$$

First matrix multiplication gives:

$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} ax_1 + dx_2 + gx_3 & bx_1 + ex_2 + hx_3 & cx_1 + fx_2 + ix_3 \end{bmatrix}$$

Second matrix multiplication gives:

$$\begin{bmatrix} ax_1 + dx_2 + gx_3 & bx_1 + ex_2 + hx_3 & cx_1 + fx_2 + ix_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$
$$= a{x_1}^2 + dx_1x_2 + gx_1x_3 + bx_1x_2 + e{x_2}^2 + hx_2x_3 + cx_1x_3 + fx_2x_3 + i{x_3}^2$$

Using the given quadratic function and the equation from the second matrix multiplication above we can solve for the variables of M:

a = 4, b = 1/2, c = 0, d = 1/2, e = 9, f = -1/2, g = 0, h = -1/2, i = 16

Therefore, the symmetric matrix for M (where M = M^T) equals to:

$$\mathbf{M} = \begin{bmatrix} \mathbf{4} & \mathbf{1/2} & \mathbf{0} \\ \mathbf{1/2} & \mathbf{9} & \mathbf{-1/2} \\ \mathbf{0} & \mathbf{-1/2} & \mathbf{16} \end{bmatrix}$$

24. *(3points) If* $M = \begin{bmatrix} 1 & 3 \\ -6 & -9 \end{bmatrix}$ *, then find M*$^{-1}$

We can use Gauss-Jordan method for solving for the inverse of M:

$[\,\mathrm{M} \mid \mathrm{I}\,] \rightarrow \text{elementary row operations} \rightarrow\ [\,\mathrm{I} \mid \mathrm{M}^{-1}\,]\,;$ where I is an identity matrix

$$\left(\begin{array}{cc|cc} 1 & 3 & 1 & 0 \\ -6 & -9 & 0 & 1 \end{array}\right) \rightarrow add\ 6 * row1\ to\ row2$$

$$\left(\begin{array}{cc|cc} 1 & 3 & 1 & 0 \\ 0 & 9 & 6 & 1 \end{array}\right) \rightarrow subtract\ 1/3 * row2\ to\ row\ 1$$

$$\left(\begin{array}{cc|cc} 1 & 0 & -1 & -1/3 \\ 0 & 9 & 6 & 1 \end{array}\right) \rightarrow divide\ row2\ by\ 9$$

$$\left(\begin{array}{cc|cc} 1 & 0 & -1 & -1/3 \\ 0 & 1 & 6/9 & 1/9 \end{array}\right)$$

$$\boldsymbol{M^{-1}} = \begin{bmatrix} \mathbf{-1} & \mathbf{-1/3} \\ \mathbf{6/9} & \mathbf{1/9} \end{bmatrix}$$

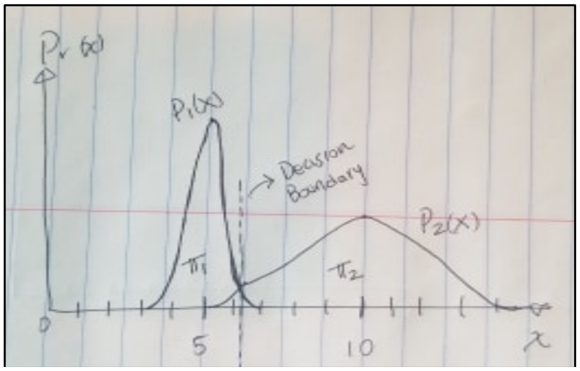
25. *Bonus: (4 points, 2 each) An experienced fisherman says that salmon length has distribution N(5,1) and sea bass length has distribution N(10,4), where N(μ,σ^2) is univariate normal distribution. The fisherman also says that prior probabilities for catching a salmon and sea bass are P(salmon) = 2/3, P(bass) = 1/3. Given this information:*

    (a) What is the decision boundary for classifying a fish caught based on its length?

Recall the univariate Gaussian function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

To find the decision boundary we first have to define each univariate Gaussian distribution, then we can equal their probabilities to each other to find the value for x where the distributions intersect (similar to the diagram below):



A few things to describe the solution. Let's first define "y" as the variable for each fish and "x" as the fish length. Then, applying Bayes rule to the probabilities we know that the evidence (denominator) is the same for each and we can cancel it out. Also, we know the mean and variance for each distribution from above, hence:

$$\Pr(y = salmon|x) = \Pr(y = bass|x)$$

$$\pi_{salmon}P_{salmon}(x) = \pi_{bass}P_{bass}(x)$$

Where,

$$P_{salmon}(x) = \frac{1}{\sqrt{2\pi*1}}\, e^{-\frac{(x-5)^2}{2*1}}\ ,\ P_{bass}(x) = \frac{1}{\sqrt{2\pi*4}}\, e^{-\frac{(x-10)^2}{2*4}}$$

$$\pi_{salmon} = \frac{2}{3}\ ,\ \pi_{bass} = 1/3$$

If we equal them with each other and solve for the algebra we get:

$$\left(\frac{2}{3}\right)\frac{1}{\sqrt{2\pi}}\, e^{-\frac{(x-5)^2}{2}} = \left(\frac{1}{3}\right)\frac{1}{\sqrt{8\pi}}\, e^{-\frac{(x-10)^2}{8}}$$

$$8\ln(4) - 4x^2 + 40x - 100 = -x^2 + 20x - 100$$

$$3x^2 - 20x - 8\ln(4) = 0$$

Using the quadratic formula we can solve for each x:

$$x_1 = -\frac{2}{3}\left(\sqrt{25+\ln(4096)}-5\right) \approx -0.51 \text{ (ignored, practically wrong)}$$

$$\boldsymbol{x_2 = \frac{2}{3}\left(\sqrt{25+\ln(4096)}+5\right) \approx 7.18} \textbf{ (This is the Decision Boundary)}$$

(b) If a fish has length 7 units, what would your prediction be based on Bayes-optimal prediction?

Probability density function of N(μ, σ^2):

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

Using Bayes-optimal prediction: $h^*(x) = \text{argmax}_j\ \pi_j P_j(x)$, where j is each label (salmon and bass). In other words, find the product for each prior with the likelihood in each label and the one with argmax is the optimal prediction.

$$\pi_{salmon}P_{salmon}(x=7) = \frac{2}{3}\frac{1}{\sqrt{2\pi * 1}}\ e^{-\frac{(7-5)^2}{2*1}} = \frac{2}{3}\frac{1}{\sqrt{2\pi}}\ e^{-2} \approx 0.03599$$

$$\pi_{bass}P_{bass}(x=7) = \frac{1}{3}\frac{1}{\sqrt{2\pi * 4}}\ e^{-\frac{(7-10)^2}{2*4}} = \frac{1}{3}\frac{1}{\sqrt{8\pi}}\ e^{-\frac{9}{8}} \approx 0.02159$$

Since the product is larger for Salmon, according to the Bayes-optimal prediction and the distributions provided for each fish is likely that a **fish with length 7 units is a Salmon.**