HW 2 Manually Graded: Upload PDF here

● Graded

Student

Emerson Liu

Total Points

30 / 30 pts

Question 1

Question 1b

4 / 4 pts

1b

Solution

To find the probability of at most adding to 9 we will calculate:

$1 - P(\text{sum of } 10) - P(\text{sum of } 11) - P(\text{sum of } 12)$

$P(\text{sum of } 12) = \frac{1}{36}$ (since there are (6)(6)=36 possible outcomes and only the outcome with 2 sixes adds to 12)

$P(\text{sum of } 11) = \frac{2}{36}$ (i.e. either 5 and 6 or 6 and 5)

$P(\text{sum of } 10) = \frac{3}{36}$ (either (6, 4), (4, 6), (5,5))

Thus $P(\text{sum of at most } 9) = 1 - \frac{3}{36} - \frac{2}{36} - \frac{1}{36} = \frac{30}{36} = \boxed{\frac{5}{6}}$

✔ **– 0 pts** Correct

Question 2

Question 1c

3 / 3 pts

1c

Solution

On any given question, you have a $1/5$ probability of guessing the correct answer and a $4/5$ probability of guessing incorrectly.

There are $C(10,3) = \frac{10!}{7!3!} = \frac{10*9*8}{3*2} = 120$ different ways to select the exact 3 questions out of 10 you get correct.

The probability of any one of these is $(\frac{1}{5})^3(\frac{4}{5})^7$.

Thus the total probability is $C(10,3)(\frac{1}{5})^3(\frac{4}{5})^7 = \boxed{120\left(\frac{4^7}{5^{10}}\right)}$

✔ **– 0 pts** Correct

Question 3

Question 2a

4 / 4 pts

2a (4 pts)

Solution:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

✔ **– 0 pts** Correct!

Question 4

Question 2b

6 / 6 pts

2b

Solution

We start by finding the value(s) of c such that $f'(c) = 0$:

$$f(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i c + \sum_{i=1}^{n} c^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2c}{n}\sum_{i=1}^{n} x_i + \frac{c^2}{n}\sum_{i=1}^{n} 1$$

Taking the derivative with respect to c we get:

$$f'(c) = 0 - \frac{2}{n}\sum_{i=1}^{n} x_i + \frac{2c}{n}(n) = -\frac{2}{n}\sum_{i=1}^{n} x_i + 2c$$

Setting this equal to 0 and solving for c:

$$f'(c) = 0 \implies -\frac{2}{n}\sum_{i=1}^{n} x_i + 2c = 0 \implies \frac{2}{n}\sum_{i=1}^{n} x_i = 2c$$

$$\implies \boxed{c = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}}$$

To test if this is a max or min, we use the 2nd derivative test:

$$f''(c) = 0 + 2$$

Thus

$$f''(\bar{x}) = 2 > 0 \implies c = \bar{x}$$

is a minimum value

✔ **– 0 pts** Correct!

Question 5

Question 3b

3 / 3 pts

SOLN: `likelihood = (p**4)*(1-p)**6`

✔ **– 0 pts** Correct

Question 6

Question 3d

4 / 4 pts

Solution:

$$\log(L(p)) = \log(p^4(1-p)^6) = 4\log(p) + 6\log(1-p).$$

$$\implies \frac{d}{dp}(\log(L(p)) = \frac{4}{p} - \frac{6}{1-p}$$

Now we solve for where the derivative equals 0:

$$\frac{d}{dp}(\log(L(p)) = 0$$

$$\implies \frac{4}{p} - \frac{6}{1-p} = 0$$

$$\implies \frac{p}{1-p} = \frac{4}{6}$$

$$\implies 6p = 4(1-p)$$

$$\implies 6p = 4 - 4p$$

$$\implies 10p = 4$$

$$\implies \boxed{\hat{p} = 0.4}$$

✔ **– 0 pts** Correct

Question 7

Question 4a

3 / 3 pts

SOLN

4ai). Represents the number of Games that the Team played in that specific year

4aii). Granularity is data about a specific team in a specific year

4aiii). Granularity is data about a specific player on a specific team in a specific year

✔ **– 0 pts** Correct

Question 8

Question 4d

3 / 3 pts

Question 4d

✔ **– 0 pts** Correct

Question assigned to the following page: 1

Question 1b) What is the probability that if I roll two 6-sided dice they add up to **at most** 9? Use LaTeX (not code) in the cell directly below to show all of your steps and fully justify your answer.

To do this problem we can find every way we can add up two dice to be greater than 9 and then subtract that probability from the total probability of rolling any sum of two dice, which is 1, to give us the probability of rolling at most 9. The ways we can roll a sum that is more than 9 is when we roll the sum of 10, 11, and 12. With these two dice we can roll a 5 on the first dice and 5 on the second, roll 5 and 6, roll 6 and 5, roll 6 and 6, roll 4 and 6, and, lastly, roll 6 and 4 to get any of these sums greater than 9. We then know that their are a total of 36 possible combinations of rolls, this is found by multiplying 6 sides by 6 sides, so therefore there is a $6/36$ chance of rolling anything greater than 9, meaning there is a $1 - 6/36 = 30/36$ chance of rolling a sum of at most 9.

No questions assigned to the following page.

Question assigned to the following page: 2

Question 1c) Suppose you uncharacteristically show up to a quiz completely unprepared. The quiz has 10 questions, each with 5 multiple choice options. You decide to guess each answer in a completely random way. What is the probability that you get exactly 3 questions correct? Use Markdown and LaTeX (not code) in the cell directly below to show all of your steps and fully justify your answer.

We can use the Bernoulli trails formula for this problem, where the number of total trais is 10, number of desired correct trails is 3, probability of success is 1/5, and probability of getting a failure is 4/5. The equation for Bernoulli trails is:

$$P(x) = \binom{n}{x} p^x * q^{n-x}$$

, where n is number of trails, x is number of desired correct trials, p is chance of success, and q is chance of failure. We can now plug our numbers into this equations:

$$P(3) = \binom{10}{3} (1/5)^3 * (4/5)^7$$

$$P(3) = 120/125 * 16384/78125$$

$$P(3) = 393216/1953125$$

No questions assigned to the following page.

Question assigned to the following page: 3

0.0.1 Question 2a)

We commonly use sigma notation to compactly write the definition of the arithmetic mean (commonly known as the average):

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The ith *deviation from average* is the difference $x_i - \bar{x}$. Prove that the sum of all these deviations is 0 that is, prove that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (write your full solution in the box directly below showing all steps and using LaTeX).

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + ... + (x_n - \bar{x})$$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 + x_2 + x_3 + ... + x_n) - n\bar{x}$$

$$= \sum_{i=1}^{n}(x_i) - n\bar{x}$$

$$= \sum_{i=1}^{n}(x_i) - \frac{n}{n}\sum_{i=1}^{n}(x_i)$$

$$= \sum_{i=1}^{n}(x_i) - \sum_{i=1}^{n}(x_i) = 0$$

No questions assigned to the following page.

Question assigned to the following page: 4

0.0.2 Question 2b)

Let $x_1, x_2, \ldots, x_n$ be a list of numbers. You can think of each index i as the label of a household, and the entry x_i as the annual income of Household i.

Consider the function

$$f(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

In this scenario, suppose that our data points $x_1, x_2, \ldots, x_n$ are fixed and that c is the only variable.

Using calculus, determine the value of c that minimizes $f(c)$. You must use calculus to justify that this is indeed a minimum, and not a maximum.

$$f(c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

$$f'(c) = \frac{-1}{n}\sum_{i=1}^{n}2(x_i - c)$$

$$= \frac{-2}{n}\sum_{i=1}^{n}(x_i - c) = \frac{-2}{n}[\sum_{i=1}^{n}(x_i) - nc]$$

$$= \frac{-2}{n}\sum_{i=1}^{n}x_i + 2c$$

$$\frac{-2}{n}\sum_{i=1}^{n}x_i + 2c = 0$$

, to find the value of c

$$2c = \frac{2}{n}\sum_{i=1}^{n}x_i$$

$$c = \frac{1}{n}\sum_{i=1}^{n}x_i$$

To find if $c = \frac{1}{n}\sum_{i=1}^{n}x_i$ is the minimum, we can take the 2nd derivative of f(c) and if it is positive then we know that the graph of f(c) is concave up at point c and consequently that c is the minimum.

$$f''(c) = \frac{d}{dc}(\frac{-2}{n}\sum_{i=1}^{n}(x_i - c)) = \frac{-2}{n}\sum_{i=1}^{n}\frac{d}{dc}(x_i - c)$$

, by the rule of summations and derivatives we can move the derivative inside of the summation.

$$= \frac{-2}{n}\sum_{i=1}^{n}(0 - 1) = \frac{-2}{n}\sum_{i=1}^{n}-1 = \frac{2}{n} * n = 2 > 0$$

Therefore since the second derivative of f(c) is postive, f(c) is concave up at the critical point of $c = \frac{1}{n}\sum_{i=1}^{n}x_i$ and is therefore the minimum.

No questions assigned to the following page.

Question assigned to the following page: 5

What is $L(p)$ (i.e. the likelihood) for the sequence TTTHTHHTTH?

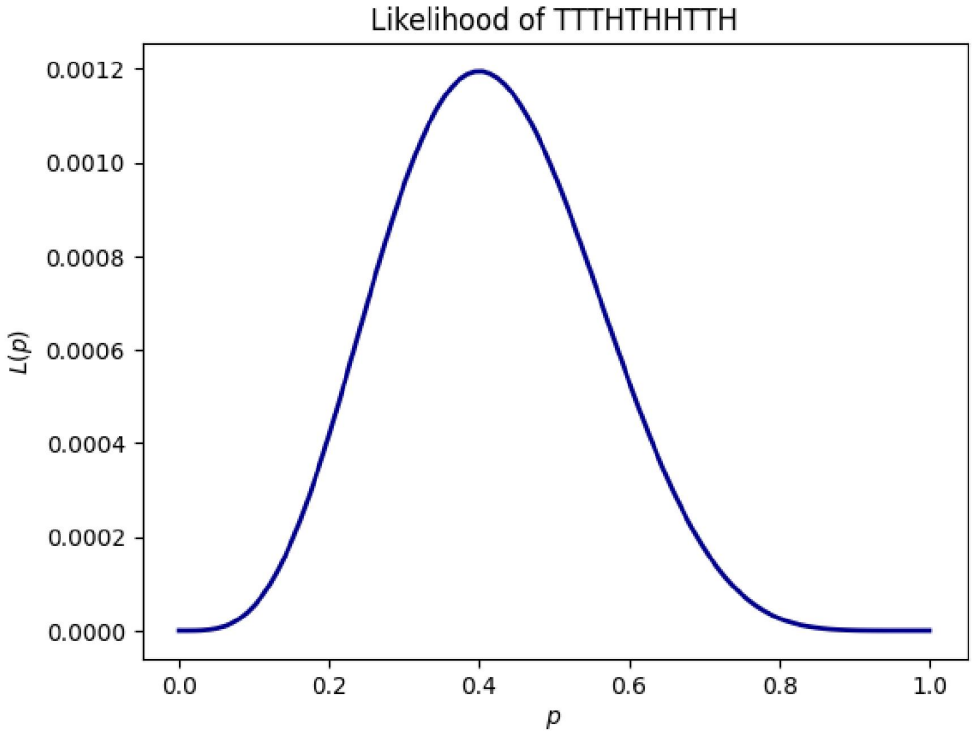
Enter your answer below by setting the `likelihood` variable equal to the correct function.

(For example `likelihood = sin(p)+2p`, althought that is definitely an incorrect answer!)

Then run the code below to plot the likelihood function.

In [211]:
```
#At the top of the notebook we already imported a useful plotting module, matplotlib with ali

p = np.linspace(0, 1, 100)
#This creates an array of 100 p-values equally spaced between 0 and 1

likelihood = p**4*(1-p)**6
#Define the likelihood function above

plt.plot(p, likelihood, lw=2, color='darkblue')
#This plots the likelihood function

plt.xlabel('$p$')
#This labels the x axis

plt.ylabel('$L(p)$')
#This labels the y-axis

plt.title('Likelihood of TTTHTHHTTH');
#This titles the plot
```

Question assigned to the following page: 5


Likelihood of TTTHTHHTTH
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
L(p)
0.0
0.2
0.4
0.6
0.8
1.0
p

Question assigned to the following page: 6

0.0.3 Question 3d)

Notice the value you found graphically for $\hat{p}$ above also intuitively makes sense because it is also the observed proportion of heads in the given sequence TTTHTHHTTH.

Let's prove what you observed graphically above. That is, let's use calculus to find $\hat{p}$.

But **wait before you start trying to find the value p where $L'(p) = 0$ (trust us, the algebra is not pretty...)**

USEFUL TIP: The value $\hat{p}$ at which the function $L(p)$ attains its maximum is the same as the value at which the function $\ln(L(p))$ attains its maximum.

This tip is hugely important in data science because many probabilities are products and the natural log function `ln` function turns products into sums. It's **much simpler to take derivatives of a sum** than a product.

Thus, to find the value p where $L'(p) = 0$: - Take the natural log `ln` of L(p) - Use properties of logs to rewrite products in `ln(L(p))` as sums - Take the derivative of this rewritten version of `ln(L(p))` - Solve `ln(L(p))=0` for p - You should get the same answer that you found graphically above.

You don't have to check that the value you've found produces a max and not a min – we'll spare you that step.

Show all steps in the cell below using Markdown and LaTeX

$$\ln(L(p)) = \ln(p^4 * (1-p)^6)$$
$$\ln(p^4 * (1-p)^6) = 4\ln(p) + 6\ln(1-p)$$
$$\frac{d}{dp}(4\ln(p) + 6\ln(1-p)) = \frac{4}{p} - \frac{6}{1-p}$$
$$\frac{4}{p} - \frac{6}{1-p} = 0$$
$$\frac{1-p}{1-p} * \frac{4}{p} - \frac{p}{p} * \frac{6}{1-p} = 0$$
$$4 - 4p - 6p = 0$$
$$10p = 4$$
$$p = 0.4$$

No questions assigned to the following page.

Question assigned to the following page: 7

0.0.4 Question 4a). EDA: Structure, Granularity and Faithfulness

Examine the structure, granularity and faithfulness of the datasets. (Hint: The common utility functions we covered in lecture will be useful here).

Then answer the following questions:

- i). What does the column `G` represent in the teams dataset? (For a description of the columns, see the documentation in the `data` folder).
- ii). What is the granularity of the `teams.csv` file?
- iii). What is the granularity of the `salary.csv` file?
- iv). How many rows and columns are in the teams dataset? Assign your answer to the variables `team_rows` and `team_col` below.
- v). How many rows and columns are in the salary dataset? Assign your answer to the variables `salary_rows` and `salary_col` below.
- vi). How many entries in the `teams.csv` file are missing Attendance Data? Assign your answer to the variable `missing_attendance` below.

0.0.5 Answer Cell for Questions 4a(i)(ii)(iii)

In this cell, answer questions 4a(i) - (iii) using Markdown (not code).

4a(i) Answer: Games

4a(ii) Answer: Team statistics and information for U.S. major league baseball teams in each specific year from 1871 to 2016

4a(iii) Answer: Player salaries and their teams for specific players from years 1985 to 2016

In the code cells below justify your answers to part (ii) and (iii) and then answer parts iv through vi

In [20]:
```
teams_df.info()

        # Show work in this cell justifying your answer to part 4a(ii) (hint: either use .value_counts
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3015 entries, 0 to 3014
Data columns (total 48 columns):
 #   Column          Non-Null Count  Dtype
```

Question assigned to the following page: 7

```
---  ------          --------------  -----
 0   yearID          3015 non-null   int64
 1   lgID            2965 non-null   object
 2   teamID          3015 non-null   object
 3   franchID        3015 non-null   object
 4   divID           1498 non-null   object
 5   Rank            3015 non-null   int64
 6   G               3015 non-null   int64
 7   Ghome           2616 non-null   float64
 8   W               3015 non-null   int64
 9   L               3015 non-null   int64
 10  DivWin          1470 non-null   object
 11  WCWin           834 non-null    object
 12  LgWin           2987 non-null   object
 13  WSWin           2658 non-null   object
 14  R               3015 non-null   int64
 15  AB              3015 non-null   int64
 16  H               3015 non-null   int64
 17  2B              3015 non-null   int64
 18  3B              3015 non-null   int64
 19  HR              3015 non-null   int64
 20  BB              3015 non-null   int64
 21  SO              2999 non-null   float64
 22  SB              2890 non-null   float64
 23  CS              2184 non-null   float64
 24  HBP             1857 non-null   float64
 25  SF              1474 non-null   float64
 26  RA              3015 non-null   int64
 27  ER              3015 non-null   int64
 28  ERA             3015 non-null   float64
 29  CG              3015 non-null   int64
 30  SHO             3015 non-null   int64
 31  SV              3015 non-null   int64
 32  IPouts          3015 non-null   int64
 33  HA              3015 non-null   int64
 34  HRA             3015 non-null   int64
 35  BBA             3015 non-null   int64
 36  SOA             3015 non-null   int64
 37  E               3015 non-null   int64
 38  DP              3015 non-null   int64
 39  FP              3015 non-null   float64
 40  name            3015 non-null   object
 41  park            2981 non-null   object
 42  attendance      2736 non-null   float64
 43  BPF             3015 non-null   int64
 44  PPF             3015 non-null   int64
 45  teamIDBR        3015 non-null   object
 46  teamIDlahman45  3015 non-null   object
 47  teamIDretro     3015 non-null   object
dtypes: float64(9), int64(26), object(13)
memory usage: 1.1+ MB
```

Question assigned to the following page: 7

In [21]:
```
salaries_df.value_counts()

# Show work in this cell justifying your answer to part 4a(iii) (hint: either use .value_count
```

Out[21]:
```
yearID  teamID  lgID  playerID   salary
1985    ATL     NL    barkele01  870000      1
2006    HOU     NL    quallch01  376000      1
        KCA     AL    brownem01  1775000     1
                      berroan01  2000000     1
                      bautide01  335500      1
                                             ..
1996    KCA     AL    lockhke01  207500      1
                      lennopa01  120000      1
                      jacomja01  150000      1
                      huismri01  118000      1
2016    WAS     NL    zimmery01  14000000    1
Name: count, Length: 26428, dtype: int64
```

In [31]:
```
# Solution Cell for 4a(iv) and 4a(v)
#Use code to find the number of rows and columns in the teams data.  Do not enter any values b

team_rows= len(teams_df)

team_col = len(teams_df.columns)

salary_rows = len(salaries_df)

salary_col = len(salaries_df.columns)
```

In [195]:
```
# Solution Cell for 4a(vi)
 # Use code to find the number of rows in the Teams data that are missing attendance data

 missing_attendance = teams_df["attendance"].isnull().sum()
```

In [196]:
```
grader.check("q4a")
```

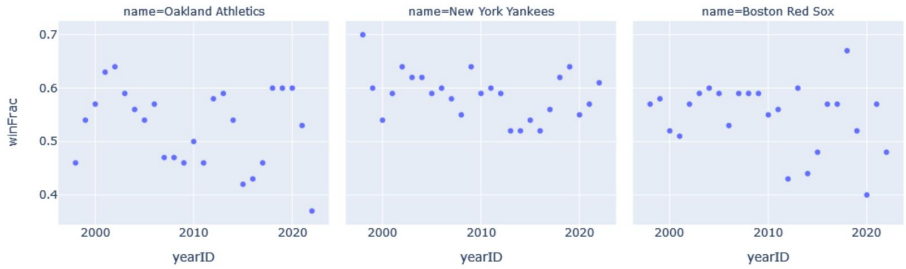
Out[196]:
```
q4a results: All test cases passed!
```

No questions assigned to the following page.

Question assigned to the following page: 8

Let's compare Oakland's winFrac to 2 other teams: Boston Red Sox and the New York Yankees:

```
In [200]: #Read the entries below to see how to create this plot:
          #Then run this cell

          px.scatter(teams_df_moneyball.loc[["OAK","NYA", "BOS"]],x="yearID", y="winFrac",facet_col = '
```



Observations: In the cell below write down at least 3 observations you can make from these plots.

1. From the first graph you can see that the Oakland A's win percentage increased from 1998 to 2002.
2. From the second graph you can observe that the Oakland A's had the worst season, in 2022, (aka the lowest win percentage) out of the A's, Yankees, and Red Sox during the years of 1998 to 2022.
3. From the second graph you can also tell who had the best season (highest win percentage) out of these three teams during the years 1998 to 2022, being the New York Yankees in 1998

No questions assigned to the following page.