Data 100, Midterm 1

Fall 2019

Name: ______________________________

Email: ______________________________@berkeley.edu

Student ID: ______________________________

Exam Room: ______________________________

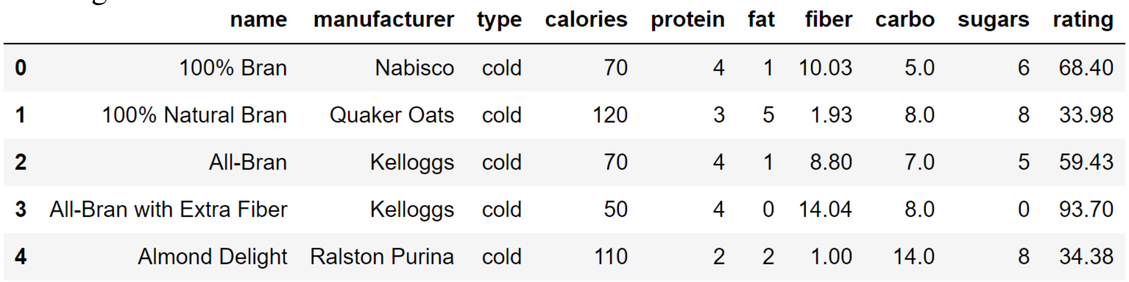
All work on this exam is my own ***(please sign)***: ______________________________

Instructions:

- This midterm exam consists of **90 points** and must be completed in the **80 minute** time period ending at **9:30**, unless you have accommodations supported by a DSP letter.
- Note that some questions have circular bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**.
- When selecting your choices, you must **fully shade** in the box/circle. Check marks will likely be mis-graded.
- You may use a one-sheet (two-sided) cheat sheet, in addition to the included Midterm 1 Reference Sheet.

1 Cereal (Pandas)

You are given a Pandas DataFrame `cereal` with information about 80 different cereals.



	name	manufacturer	type	calories	protein	fat	fiber	carbo	sugars	rating
0	100% Bran	Nabisco	cold	70	4	1	10.03	5.0	6	68.40
1	100% Natural Bran	Quaker Oats	cold	120	3	5	1.93	8.0	8	33.98
2	All-Bran	Kelloggs	cold	70	4	1	8.80	7.0	5	59.43
3	All-Bran with Extra Fiber	Kelloggs	cold	50	4	0	14.04	8.0	0	93.70
4	Almond Delight	Ralston Purina	cold	110	2	2	1.00	14.0	8	34.38

The figure above is the result of running `cereal.head()`. All values are per-serving. `type` can be either `cold` or `hot`. `rating` is the average score (out of 100) given by customers.

(a) [1 Pt] What is the granularity of the cereal data frame?

○ serving of cereal

○ type of cereal

○ manufacturer

○ **name of cereal**

(b) [3 Pts] Add a new column to `cereal` named `low_calorie` which has the boolean value `True` if the cereal is low-calorie and `False` otherwise. A cereal is low-calorie if it has less than or equal to 100 calories per serving.

`cereal[`____________________________`] =` ________________________________

Solution:

```
cereal["low_calorie"] = cereal["calories"] <= 100
```

(c) [3 Pts] Identify the type for each of the following variables.

`fiber`

○ **continuous** ○ discrete ○ nominal ○ ordinal

`type`

○ continuous ○ discrete ○ **nominal** ○ ordinal

`low_calorie`

○ continuous ○ discrete ○ nominal ○ **ordinal**

(d) [6 Pts] For this problem and the next problem below you may use the following functions:

```
groupby, agg, filter, merge
unique, value_counts, sort_values, apply
max, min, mean, median, std, count,
np.mean, sum, any, all, isnull, len
```

You can also use any other methods you have used in class, for example, anything in the Pandas or Numpy libraries. **You can leave lines inside parentheses blank to represent a function call with no arguments**.

Create a Series indexed by manufacturer. For each manufacturer, the value should be equal to the maximum `sugars` value of all cereals by that manufacturer. Your series should be sorted by the value in decreasing order. You may not need all lines.

```
max_sugar = cereal.____________________(______________________)["sugars"]
                  .____________________(______________________)
                  .____________________(______________________)
```

For example, the first few entries of the Series would be:

```
Kelloggs 15
Post 14
Quaker Oats 14
```

Solution:

```
cereal.groupby("manufacturer")["sugars"] \
     .agg(max).sort_values(ascending=False)
```

(e) [8 Pts] Which manufacturers make **only** cold cereals? Return an array, list, or series of these manufacturers that exclusively manufacture cold cereals, i.e. your list should not a company if it makes ANY hot cereals. You may not need all lines. **You can leave lines inside parentheses blank to represent a function call with no arguments**.

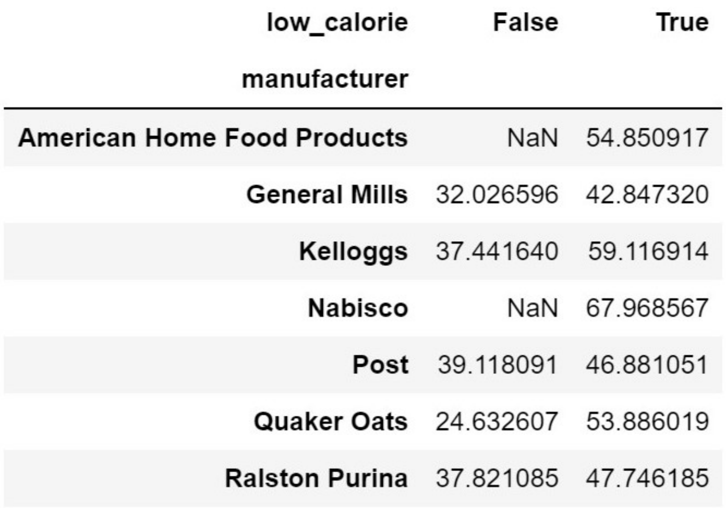
```
def f(df):
    ______________________________________________________
    ______________________________________________________
    ______________________________________________________

cold_only =
      cereal.____________________(______________________)
            .____________________(______________________)["manufacturer"]
            .____________________(______________________)
```

Solution:

```
cereal.groupby("manufacturer") \
    .filter(lambda x: sum(x["type"] == "hot") == 0)["manufacturer"]
```

(f) [2 Pts] Consider the data frame below. Assume `cereal` was modified correctly in part a. The Interior values are average rating per category, e.g. 32.026596 is the average rating of the low calorie cereals made by General Mills.



low_calorie	False	True
manufacturer		
American Home Food Products	NaN	54.850917
General Mills	32.026596	42.847320
Kelloggs	37.441640	59.116914
Nabisco	NaN	67.968567
Post	39.118091	46.881051
Quaker Oats	24.632607	53.886019
Ralston Purina	37.821085	47.746185

Which of the following four lines of code could be used to create this data frame?

□ **`pd.pivot_table(data=cereal, index=’manufacturer’, columns=’low_calorie’, values=’rating’, aggfunc=np.mean)`**

□ `cereal.groupby([’manufacturer’,’low_calorie’])[’rating’].mean()`

□ `pd.pivot_table(data=cereal, index=’low_calorie’, columns=’manufacturer’, values=’rating’, aggfunc=np.mean)`

□ `cereal.groupby(’rating’)[[’manufacturer’,’low_calorie’]].mean()`

Solution:

(g) [2 Pts] The above table contains NaNs because some companies don’t make cereals with the given calorie level. By calorie level, we mean whether `low_calorie` is True or False. E.g. Nabisco does not make any low calorie cereals. If we wanted to show a colleague this pivot table to illustrate the average rating across manufacturer and calorie level combinations for cereals, what should we do with these NaN values? Pick the one best option that applies.

○ Fill them with the average rating across all the cereals in the same calorie level

○ Fill them with the average rating across all cereals from the same manufacturer

○ Both A and B are acceptable

○ **Leave as-is**

○ Replace with a rating randomly selected from a cereal with the same calorie level

Solution: You want to leave the values as-is because it will show that there are no such cereals in the specified category which is useful information in itself.

2 Computing Summary Statistics

Suppose we're given the set of points $\{-15, 10, 20, 30, 30, 35, 40, 50\}$, and we want to determine a summary statistic c. For each of the following loss functions, determine or select the optimal value of c, $\hat{c}$, that minimizes the corresponding empirical risk.

For (a), (b), and (c), select the correct answer. For (d), write your answer in the provided box.

To help you with this task, we have computed the following: mean = 25, median = 30, SD = 20, and $n = 8$.

(a) [2 Pts] $(x_i - c)^2$

○ 20
○ **25**
○ 5
○ 50
○ 0

(b) [2 Pts] $5(x_i - c)^2$

○ 20
○ 125
○ **25**
○ 100
○ -15

(c) [2 Pts] $|x_i - c|$

○ 25
○ 50
○ 40
○ **30**
○ -15

(d) [5 Pts] $(3x_i - c)^2$

For part d, write only your answer in the box below. Feel free to show your work elsewhere on this page.

$\hat{c}$ =

Solution: $\hat{c} = 3\bar{x} = 75$

3 Regex

You are interviewing for a job at Triple Rock, and they want you to prove your skills with regular expressions on synthetic data.

(a) [4 Pts] First they give you a list of two of their distributors.

```
distributors[0] = "Geyser Beverage:  55 Wright Brothers Ave",
distributors[1] = "Mindful Distribution:  2935 Adeline St"
```

Give a regular expression that extracts the name and street number from such strings. **Your regex should work for either string**. For example, after running the code below, `name` should be `'Geyser Beverage'` and `street_number` should be `55`.

```
regex_1 = r' ________________________________________________'
name, street_number = re.findall(regex_1, distributors[0])[0]
```

Solution:

```
([\w\s]+):\s+(\d+)
```

(b) Next they give you information regarding how each of two table paid its bill. The first two entries are:

```
paid_bills[0] = "4123713131673827 paid $30.50,
and $37 paid by 5612512165638672.",

paid_bills[1] = "$171.25 was charged to 4612512165638672."
```

i. [4 Pts] Give a regular expression that can extracts Visa and Mastercard credit card numbers from such strings. **Your regex should work for either string**. A Visa credit card number is any sequence of 16 digits that starts with a 4, and a Mastercard is any sequence of 15 digits that starts with 5. Observe there are no dashes or other extraneous characters in a credit card number. For example, after running the code below, `cc_nums` should be `['4123713131673827', '5612512165638672']`.

```
regex_2 = r' ________________________________________________'
cc_nums = re.findall(regex_2, paid_bills[0])
```

Solution:

```
[45]\d{15}
```

ii. [4 Pts] Write a regular expression which will correctly find and return the dollar amounts including whatever is to the right of the optional decimal point. **Your regex should work for either string**. For example, after running the code below, `amounts` should be `['30.50', '37']`.

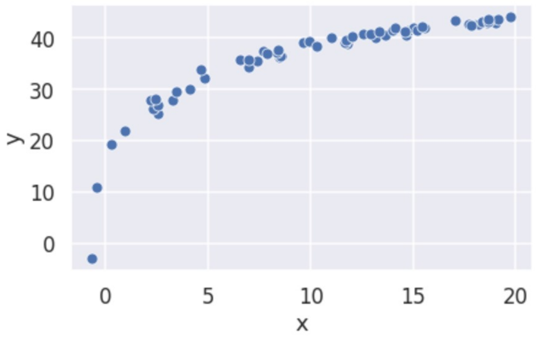
```
regex_3 = r' ____________________________________________'
amounts = re.findall(regex_3, paid_bills[0])
```

Solution:

```
\$(\d+\.?\d*)
```

4 EDA

(a) [2 Pts] Which of the following transformations could help make linear the relationship shown in the plot below? Select all that apply:



□ $\log(y)$

□ $\log(x)$

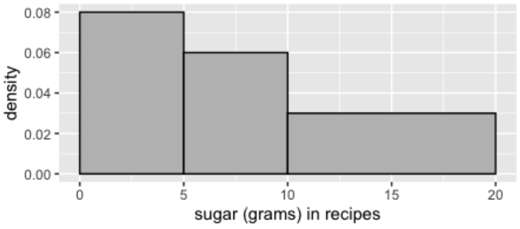
□ e^x

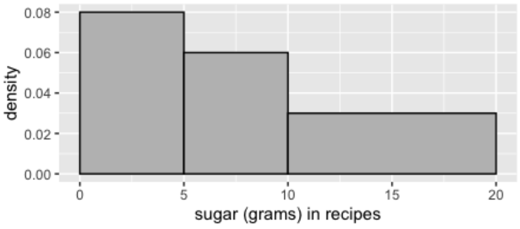
□ y^3

□ None of the Above

(b) Sally likes making desserts, and she wants to learn more about the sugar content of her recipes. For each of 100 recipes, she records the amount of sugar (in grams) per serving. A histogram of the sugar measurements appears in the plot below on the left.

Her friend, Max, makes a similar histogram of his 100 recipes, which appears below on the right. Note: The two images shown are exactly alike.





i. [3 Pts] How many of Sally's recipes have more than 10 grams of sugar per serving? Do not worry about interval endpoints, i.e. assume that no recipes have exactly 0, 5, 10, or 20 grams.

○ 15 ○ **30** ○ 60 ○ Impossible to tell

ii. [3 Pts] How would you describe the distribution of sugar in Sally's recipes? Check all that apply.

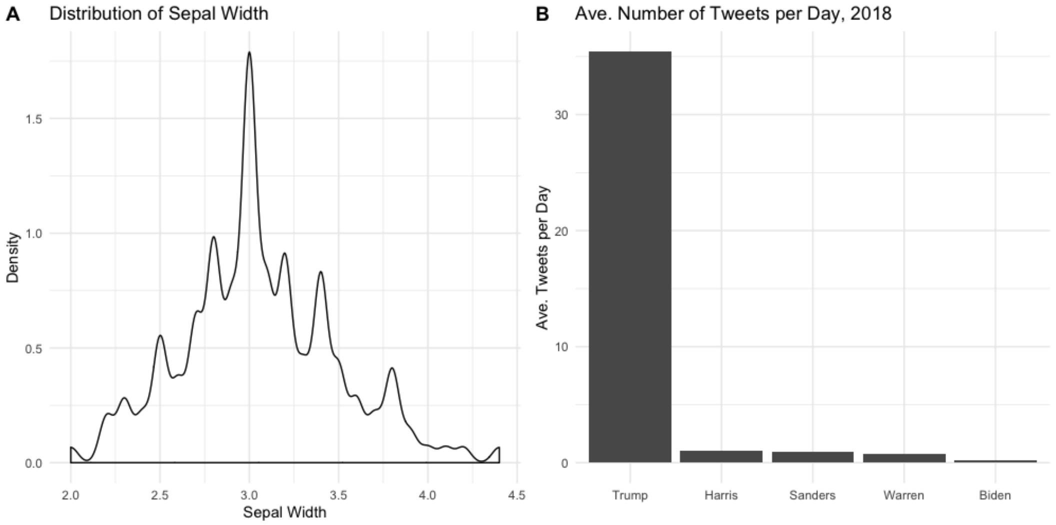
□ **unimodal** □ multimodal □ symmetric
□ skew left □ **skew right** □ contains outliers

iii. [2 Pts] Do Max's recipes have the exact same sugar content as Sally's?

○ Yes ○ No ○ **Impossible to tell**

5 Visualizations

(a) [6 Pts] Consider plots A and B below. For each plot, identify its **primary flaw** (if any) and give a recommendation in the provided box.



Does Plot A have a significant flaw? ○ **Yes** ○ No

If you picked yes, in the box below, make a recommendation to fix the primary flaw.

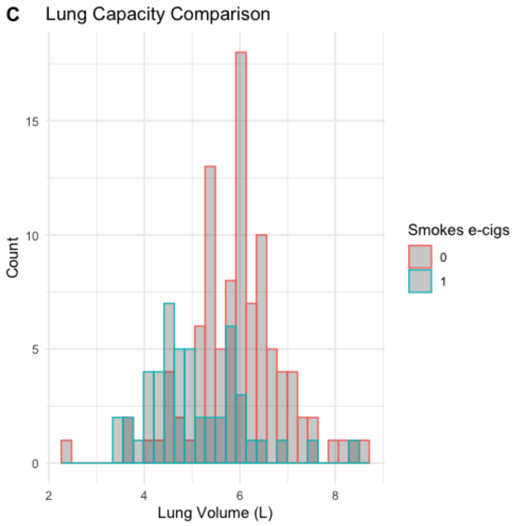
Solution: Increase the bandwidth for a smoother density estimate

Does Plot B have a significant flaw? ○ **Yes** ○ No

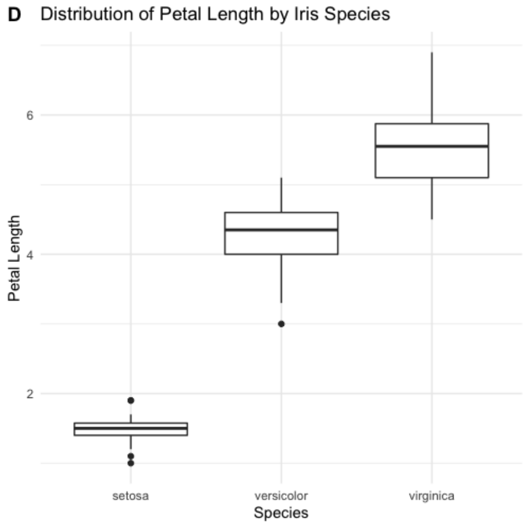
If you picked yes, in the box below, make a recommendation to fix the primary flaw.

Solution: Re-scale y-axis

(b) [6 Pts] Consider plots C and D below. For each plot, identify its **primary flaw** (if any) and give a recommendation in the provided box.



C Lung Capacity Comparison



D Distribution of Petal Length by Iris Species

Does Plot C have a significant flaw? ○ **Yes** ○ No

If you picked yes, in the box below, make a recommendation to fix the primary flaw.

Solution: Make density curves so that you can more easily compare the distributions.

Does Plot D have a significant flaw? ○ Yes ○ **No**

If you picked yes, in the box below, make a recommendation to fix the primary flaw.

Solution: Not applicable

(c) [2 Pts] For plot **D** above, which species of Iris has the highest frequency in the dataset?

○ Virginica
○ Versicolor
○ Setosa
○ **Impossible to tell**

6 Sampling

(a) Professor Hug is an instructor for both Data 100 and CS W186 this semester. Students come to his office hours for both of these classes, but some students also come for other reasons. Professor Hug is interested in knowing how many Data 100 students this semester have taken CS 61B. He takes a convenience sample of people that come to office hours.

i. [2 Pts] Name a group or individual that is included in the sampling frame but is not in the population of interest.

Solution: CS W186 Students who are not in Data 100; students not in Data 100 who come to Professor Hug's office hours for other reasons not related to Data 100 and CS W186

ii. [2 Pts] Name a group or individual that is in the population of interest but not in the sampling frame.

Solution: Data 100 students who do not go to Professor Hug's office hours

(b) For the rest of the question, assume that the sampling frame is the exact same as the population. That is, both the population and the sampling frame is the population of Data 100 students. Also, assume that there are 1000 students in Data 100 and that 500 of them have taken CS 61B. Using the class list, Professor Hug takes a simple random sample of 50 students in Data 100. Let X_i be 1 if the i^{th} person sampled took CS 61B and 0 otherwise, for $i = 1, \dots, 50$.

Find the following quantities. Somewhere in this problem you might need the "finite population correction factor" given by $\frac{N-n}{N-1}$.

i. [3 Pts] $P(X_5 = 1) =$ □

Solution: $P(X_5 = 1) = P(\text{5th person has taken CS 61B}) = \frac{500}{1000} = \frac{1}{2}$

ii. [4 Pts] $P(X_5 = 1, X_{50} = 1) =$ □

Solution: $P(X_5 = 1, X_{50} = 1) == \frac{500}{1000} \times \frac{499}{999} = \frac{1}{2} \times \frac{499}{999}$

iii. [3 Pts] $Var(X_1 + X_2 + \cdots + X_{50}) =$ □

Solution:

$$\begin{aligned} Var(X_1 + X_2 + \cdots + X_{50}) &= \frac{N-n}{N-1} np(1-p) \\ &= \frac{1000-50}{1000-1} 50 \frac{500}{1000}(1 - \frac{500}{1000}) \\ &= \frac{950}{999} 50 \frac{1}{4} \end{aligned}$$

(c) [4 Pts] Now suppose Professor Hug takes a census of the class. Let X_i be 1 if the i^{th} person sampled took CS 61B and 0 otherwise, for $i = 1, \dots, 1000$. Find the given quantity.

$Var(X_1 + X_2 + \cdots + X_{1000}) =$ □

Solution: If a census is taken, then all 1000 students are surveyed, and $X_1 + \cdots + X_{1000} = 500$. That is, there is no variability in this sum, and

$$Var(X_1 + X_2 + \cdots + X_{1000}) = 0$$

Note that the finite population correction factor is 0 in this case.

Solution: $\frac{1}{2}$. Once we know one of the values (say X_{4996}), we know that the rest of the values must be the same since there is only one person left to sample from. Since $P(X_{4996} = 1) = \frac{1}{2}$ we know that $P(X_{4996} = 1, X_{4997} = 1, X_{4998} = 1, X_{4999} = 1, X_{5000} = 1) = \frac{1}{2}$.

For the rest of the question, assume that the sampling frame is the exact same as the population. That is, both the population and the sampling frame is the population of Data 100 students. Assume that there are 1000 students that are taking Data 100 and that 500 of them have taken CS 61B with Professor Hug in the past.

(d) Now, assume that no two people come to office hours at the exact same time, so there is an ordering of people who come to office hours. Every fifth person who comes to Professor Hug's office hours decides that his office hours aren't useful, so they don't come back anymore.

On the day before the midterm, a total of 5000 students show up to his office hours. Let X_i be an indicator random variable representing whether the i^{th} student has taken CS 61B with Professor Hug in the past. That is $X_i = 1$ if the i^{th} student has taken CS 61B with him, and $X_i = 0$ if the student has not. For the rest of the questions, you can leave your answer as a product of fractions (or a fraction to an exponent).

(e) What is $P(X_{4996} = 1, X_{4997} = 1, X_{4998} = 1, X_{4999} = 1, X_{5000} = 1)$?

Solution: $\frac{1}{2}$. Once we know one of the values (say X_{4996}), we know that the rest of the values must be the same since there is only one person left to sample from. Since $P(X_{4996} = 1) = \frac{1}{2}$ we know that $P(X_{4996} = 1, X_{4997} = 1, X_{4998} = 1, X_{4999} = 1, X_{5000} = 1) = \frac{1}{2}$.

(f) What is $P(X_5 = 1, X_{10} = 1, X_{15} = 1)$?

Solution: $\frac{500}{1000} * \frac{499}{999} * \frac{498}{998}$. Once we select an individual for one of these random variables, we cannot select them again.

(g) What is $\mathbb{E}[\sum_{i=1}^{5000} X_i] = \mathbb{E}[X_1 + X_2 + X_3 + \cdots X_{5000}]$?

Solution: $\mathbb{E}[\sum_{i=1}^{5000} X_i] = 5000\mathbb{E}[X_1] = 2500$.

(h) What is Var$(X_5 + X_{10} + X_{15} + X_{20})$?

Solution: This is the same as taking an SRS of 4 people from the population of 1000. Hence, Var$(X_5 + X_{10} + X_{15} + X_{20}) = \frac{1000-4}{1000-1} * Var(X_5) = \frac{996}{999} * \frac{1}{2} * \frac{1}{2} = \frac{249}{999}$.

(i) What is Var$(\sum_{i=1}^{1000} X_{5i}) =$ Var$(X_5 + X_{10} + X_{15} + \cdots X_{5000})$?

Solution: If 5000 people come to his office hours, then we know that all 1000 students come at some point since every 5th person leaves. Note that $X_5 + X_{10} + X_{15} + \cdots X_{5000} = 500$ since the sum of these random variables represent how many people in Data 100 have taken CS 61B with Professor Hug. Hence $Var(X_5 + X_{10} + X_{15} + \cdots X_{5000}) = Var(500) = 0$.

(j) What is Var$(X_1 + X_2 + X_3 + X_4 + \sum_{i=1}^{1000} X_{5i}) =$ Var$(X_1 + X_2 + X_3 + X_4 + X_5 + X_{10} + X_{15} + \cdots X_{5000})$?

Solution: Note that $X_1 + X_2 + X_3 + X_4$ is independent from $\sum_{i=1}^{1000} X_{5i}$ and X_1 through X_4 are independent. Hence, Var$(X_1 + X_2 + X_3 + X_4 + \sum_{i=1}^{1000} X_{5i}) =$ Var$(X_1 + X_2 + X_3 + X_4) +$ Var$(\sum_{i=1}^{1000} X_{5i}) =$ Var$(X_1 + X_2 + X_3 + X_4) + 0 = 4Var(X_1) = 4 * \frac{1}{2} * \frac{1}{2} = 1$.