Discussion 1

Disc103

Probability/Pandas

SCAN ME

Attendance form

Password: rube

1. Course/Discussion Logistics

Announcements

Due Dates

Homework 1 - June 27

Lab 1 - June 27

Lab 2 - June 27

Other

Monday/Thursday OH most crowded

Contact: mko357@berkeley.edu

Discussion

Zoom- Tues/Thurs 1-2pm

New password/form each week

3 drops

Attendance → 10% of exam

2. Introductions

Name

Major/Year

something you're excited for this summer discussion 0

review of prereqs

3. Worksheet

^ Distribution

Binomial Formula example

iid

$$P(X = k) = \overbrace{\underbrace{\binom{n}{k}} p^k (1-p)^{n-k}}^{\text{Probability Mass Function for a Binomial}}$$

Probability that our variable takes on the value k

n - total sample size

(10)

p - probability of heads (.5)

k - # of successes (3)

$$\binom{10}{3}(.5)^3(.5)^7$$

$$\underline{H}\ \ \underline{H}\ \ \underline{H}\ \ \underline{T}\ \ \underline{T}\ \ \underline{T}\ \ \underline{T}\ \ \underline{T}\ \ \underline{T}\ \ \underline{T}$$

can choose multiple answers

1. Consider a sample of size n where n is a positive integer drawn at random with replacement from a population in which a proportion p of the individuals are called successes.

   (a) For an integer k such that $0 \leq k \leq n$, which of the following are equal to the chance of getting exactly k successes in the sample?

   $$\binom{n}{k} = \binom{n}{n-k} = \frac{n!}{k!(n-k)!}$$

   (i) $p^k(1-p)^{n-k}$

   [✓ (ii)] $\binom{n}{k}p^k(1-p)^{n-k}$

   [✓ (iii)] $\binom{n}{n-k}p^k(1-p)^{n-k}$

   [✓ (iv)] $\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$

   (b) Which of the following are equal to the chance of getting at least one success in the sample?

   (i) $np(1-p)^{n-1}$

   (ii) $\sum_{k=2}^{n} \binom{n}{k}p^k(1-p)^{n-k}$

   [✓ (iii)] $\sum_{k=1}^{n} \binom{n}{k}p^k(1-p)^{n-k}$

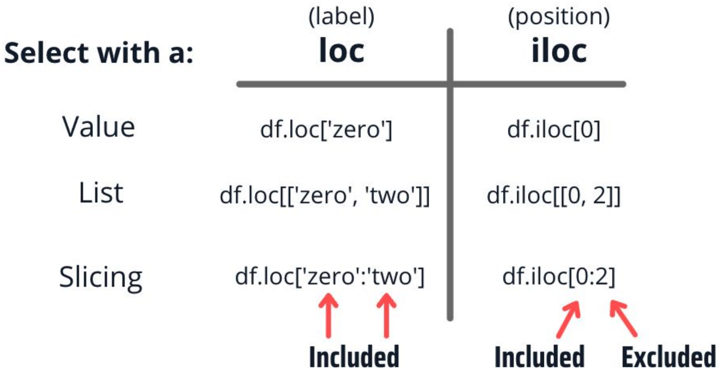
   (iv) $1 - p^n$

   [✓ (v)] $1 - (1-p)^n$

Important Pandas commands/operations

iloc vs. loc: iloc = "integer" / "index" location

loc selects rows and columns with specific labels. iloc selects rows and columns at specific integer positions

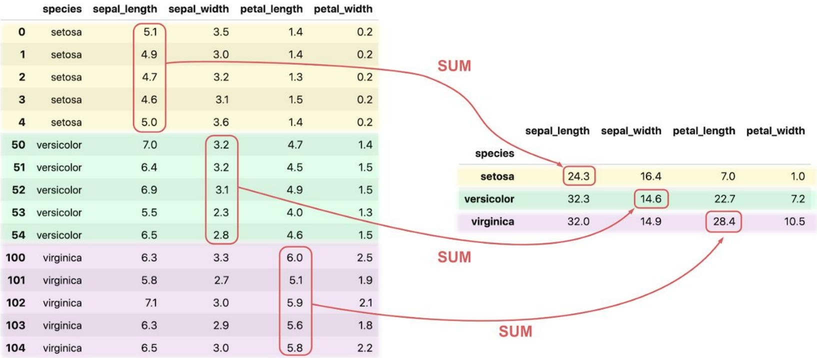


Select with a:	(label) **loc**	(position) **iloc**
Value	df.loc['zero']	df.iloc[0]
List	df.loc[['zero', 'two']]	df.iloc[[0, 2]]
Slicing	df.loc['zero':'two']	df.iloc[0:2]
	Included	**Included** **Excluded**

Important Pandas commands/operations

groupby:

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

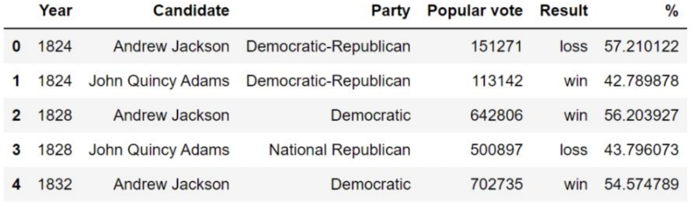


	species	sepal_length	sepal_width	petal_length	petal_width
0	setosa	5.1	3.5	1.4	0.2
1	setosa	4.9	3.0	1.4	0.2
2	setosa	4.7	3.2	1.3	0.2
3	setosa	4.6	3.1	1.5	0.2
4	setosa	5.0	3.6	1.4	0.2
50	versicolor	7.0	3.2	4.7	1.4
51	versicolor	6.4	3.2	4.5	1.5
52	versicolor	6.9	3.1	4.9	1.5
53	versicolor	5.5	2.3	4.0	1.3
54	versicolor	6.5	2.8	4.6	1.5
100	virginica	6.3	3.3	6.0	2.5
101	virginica	5.8	2.7	5.1	1.9
102	virginica	7.1	3.0	5.9	2.1
103	virginica	6.3	2.9	5.6	1.8
104	virginica	6.5	3.0	5.8	2.2

	sepal_length	sepal_width	petal_length	petal_width
species				
setosa	24.3	16.4	7.0	1.0
versicolor	32.3	14.6	22.7	7.2
virginica	32.0	14.9	28.4	10.5

Pandas Practice

Below are the first few rows of the `elections` DataFrame from lecture.



	Year	Candidate	Party	Popular vote	Result	%
0	1824	Andrew Jackson	Democratic-Republican	151271	loss	57.210122
1	1824	John Quincy Adams	Democratic-Republican	113142	win	42.789878
2	1828	Andrew Jackson	Democratic	642806	win	56.203927
3	1828	John Quincy Adams	National Republican	500897	loss	43.796073
4	1832	Andrew Jackson	Democratic	702735	win	54.574789

5. We want to select the ”Popular vote” column as a `pd.Series`. Which of the following lines of code will error?

   A) `elections[‘Popular vote’]` ✓

   [✓ B)] `elections.iloc[‘Popular vote’]` ✗

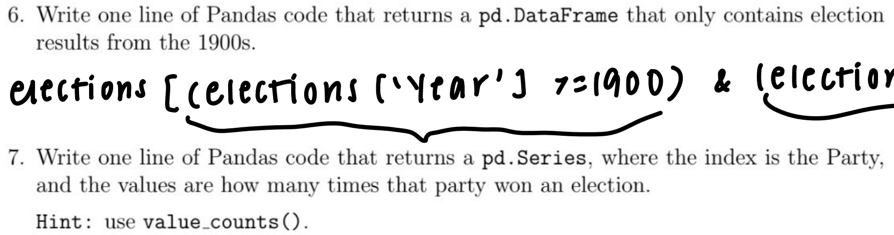
   [✓ C)] `elections.loc[‘Popular vote’]` ✗

   D) `elections.loc[:, ‘Popular vote’]`

   [✓ E)] `elections.iloc[:, ‘Popular vote’]` ✗

: = all

.loc [row(s), column(s)]



6. Write one line of Pandas code that returns a pd.DataFrame that only contains election results from the 1900s.

```
elections [(elections ['Year'] >=1900) & (elections ['Year'] <2000)]
```

7. Write one line of Pandas code that returns a pd.Series, where the index is the Party, and the values are how many times that party won an election.

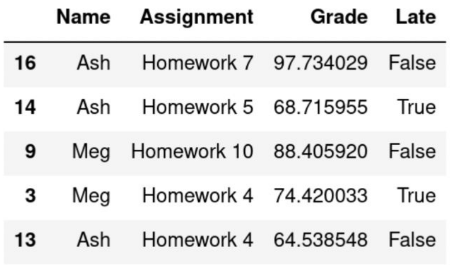
Hint: use value_counts().

```
elections [elections [`Result'] == `win'][`Party'].value_counts()
```

created a table of wins

8. Anirudhan is writing a grading script to compute grades for students in Data 101. Recall that many factors go into computing a student's final grade, including homework, discussion, exams, and labs. In this question, we will help Anirudhan compute the homework grades for all students using a DataFrame, `hw_grades`, provided by Gradescope.

The Pandas DataFrame `hw_grades` contains homework grades for all students for all homework assignments, with one row for each combination of student and homework assignment. **Any assignments that are incomplete are denoted by NaN (missing) values, and any late assignments are denoted by a True boolean value in the Late column.** You may assume that the names of students are unique. Below is a sample of `hw_grades`.



	Name	Assignment	Grade	Late
16	Ash	Homework 7	97.734029	False
14	Ash	Homework 5	68.715955	True
9	Meg	Homework 10	88.405920	False
3	Meg	Homework 4	74.420033	True
13	Ash	Homework 4	64.538548	False

(a) Find the total number of late homework submissions.

(b) Find Meg's average homework grade. Assume there are no late penalties.