Data 100, Spring 2022 **January 28, 2021**

Discussion #2

Finding Chances

Golden rules for finding the probability of an event:

- Addition Rule: list all the distinct ways the event can happen, and add the chances of all the ways. Note that all events must be mutually exclusive for this rule to apply!
- Complement Rule: if the list above looks long and complicated, make the list of ways in which the event *doesn't* happen and calculate its probability q; it might be simpler. The probability of the original event happening p is the complement of q: $p = 1 - q$.
- Multiplication Rule: If an event involves multiple independent trials, like a number of random draws, imagine yourself conducting the experiment one trial at a time. The probability of the event is the product of the probabilities of each trial.

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?

   (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) $1 - (1-p)^n$

Sampling and Bias

2. It's time for the Data 100 midterm and the professor wants to estimate the difficulty of the exam. They decide to survey students on the exam's difficulty with a 10-point scale and then use the mean of the students' responses as the estimate.

   (a) What is the population the professor is interested in trying to understand?

   ○ A. Students in Data 100
   ○ B. Students enrolled in the Data 100 Piazza
   ○ C. Students who attend the Data 100 lecture
   ○ D. Students who took the Data 100 midterm

   (b) The professor considers a few different methods for collecting the survey data. Which of the following methods is best? (think through which considerations go into "best")

   ○ A. The professor sends a Zoom poll to all students in an optional midterm debrief lecture after going over exam solutions.
   ○ B. The professor adds a question to the homework assignments of a simple random sample of students within every discussion section.
   ○ C. The professor makes a post on Piazza asking students to submit an anonymous Google Form containing the survey question.
   ○ D. The professor chooses a simple random sample of discussion sections, goes to each selected section, and asks each student in the group as part of the final discussion question.

3. A campus organization wants to take a sample of Berkeley students who are registered for classes this semester. To do this, the organization takes a simple random sample of 20 classes from among all classes offered this semester, and then takes all students in those classes. You can assume that the organization has access to complete enrollment information all classes.

   (a) Is this a simple random sample of students? Explain.

   (b) Is this a probability sample of students? Explain.

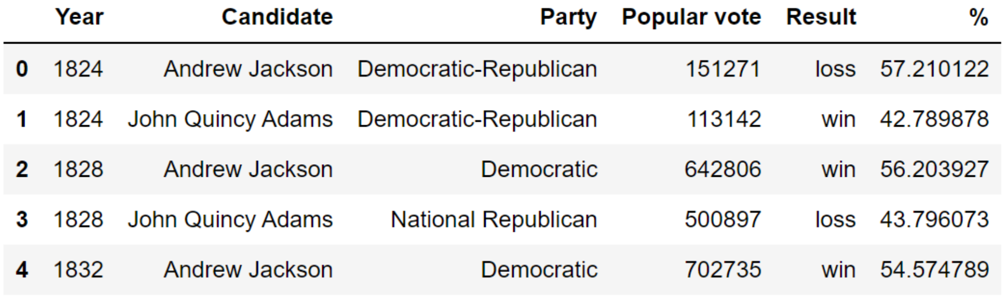
4. The Current Population Survey is a national survey run by the Census Bureau. It is thorough and reliable, and thus is sometimes used as a benchmark to assess the accuracy of other surveys.

   As part of an assessment of its own phone surveys, the Pew Research Center found that the response rates have been dropping over the years. Still, on most measures, its estimates were comparable to those of the Current Population Survey. For example, 55% of respondents in the most recent Pew Survey said they did some type of volunteer work for or through an organization in the past year in a phone survey, compared to 27% in the Current Population Survey.

How do you think this difference might have arisen?

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']`

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

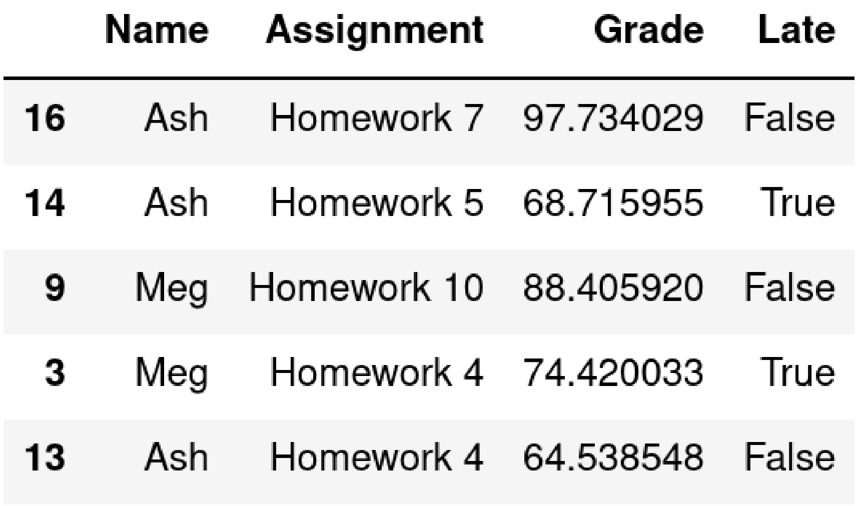
7. Write one line of Pandas code that returns a `pd.Series`, where the index is the Party, and the values are how many times that party won an election.

   `Hint:` use `value_counts()`.

Grading Assistance (Bonus)

8. Fernando 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 Fernando 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) Assuming there is a late penalty that causes a 10% grade reduction to the student's current score (i.e. a 65% score would become a 65% - 6.5% = 58.5%), write a line of Pandas code to calculate all the homework grades, including the late penalty if applicable, and store it in a column named `'LPGrade'`.

   (b) Which of the following expressions outputs the students' names and number of late assignments, from least to greatest number of late assignments?

   ○ A. `hw_grades.groupby(['Name']).sum().sort_values()`

   ○ B. `hw_grades.groupby(['Name', 'Late']).sum().sort_values()`

○ C. `hw_grades.groupby(['Name']).sum()['Late'].sort_values()`

○ D. `hw_grades.groupby(['Name']).sum().sort_values()['Late']`

(c) If each assignment is weighted equally, fill in the blanks below to calculate each student's overall homework grade, including late penalties for any applicable assignments.

Hint: Recall that incomplete assignments have NaN values. How can we use `fillna` to replace these null values?

```
hw_grades.______________(______________) \
          .groupby(______________)[______________] \
          .agg(______________)
```

(d) Of all the homework assignments, which are the most difficult in terms of the median grade? Order by the median grade, from lowest to greatest. Do not consider incomplete assignments or late penalties in this calculation.

Fill in the blanks below to answer this question.

Hint: Recall that incomplete assignments have NaN values. How can we use `dropna` to remove these null values?

```
hw_grades.______________() \
          .groupby(______________)[______________] \
          .agg(______________) \
          .sort_values()
```