Exercises

1. Consider the following text data describing purchases of financial products:

Id	Date	Product	Company
0	99/99/99	Debt collection	California Accounts Service
1	06/15/10	Credit reporting	EXPERIAN INFORMATION SOLUTIONS INC
3	10/21/14	MORTGAGE	OCWEN LOAN SERVICING LLC
5	03/30/15		The CBE Group Inc
6	02/03/16	Debt collection	The CBE Group, Inc.
7	01/07/17	Credit reporting	Experian Information Solutions Inc.
8	03/15/17	Credit card	FIRST NATIONAL BANK OF OMAHA

(1) [2 Pts] Select all the true statements from the following list.

- ☐ Some of the product values appear to be missing.
- ☐ Some of the date values appear to be missing.
- ☐ The file is comma delimited
- ☐ The file is fixed width formatted.
- ☐ To analyze the companies we will need to correct for variation in capitalization and punctuation.
- ☐ None of the above statements are true.

1. Consider the following text data describing purchases of financial products:

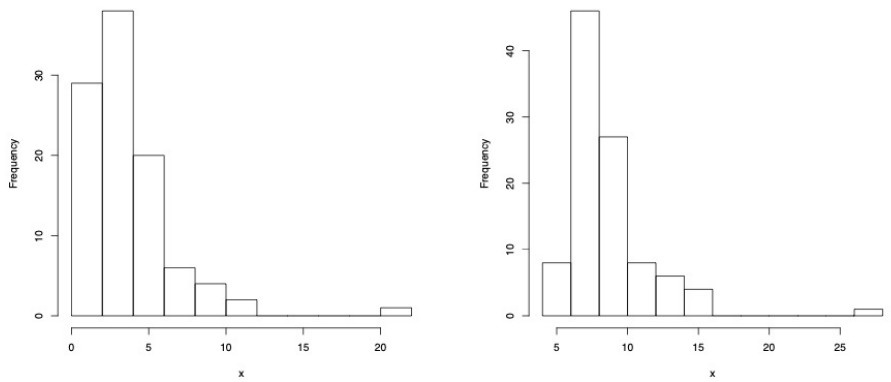
Id	Date	Product	Company
0	99/99/99	Debt collection	California Accounts Service
1	06/15/10	Credit reporting	EXPERIAN INFORMATION SOLUTIONS INC
3	10/21/14	MORTGAGE	OCWEN LOAN SERVICING LLC
5	03/30/15		The CBE Group Inc
6	02/03/16	Debt collection	The CBE Group, Inc.
7	01/07/17	Credit reporting	Experian Information Solutions Inc.
8	03/15/17	Credit card	FIRST NATIONAL BANK OF OMAHA

(1) [2 Pts] Select all the true statements from the following list.

- √ **Some of the product values appear to be missing.**
- √ **Some of the date values appear to be missing.**
- ☐ The file is comma delimited
- √ **The file is fixed width formatted.**
- √ **To analyze the companies we will need to correct for variation in capitalization and punctuation.**
- ☐ None of the above statements are true.

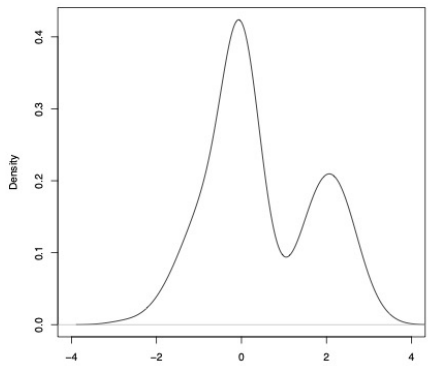
(a) (2 pt) Are the two histograms below displaying exactly the same data? Circle only one answer.

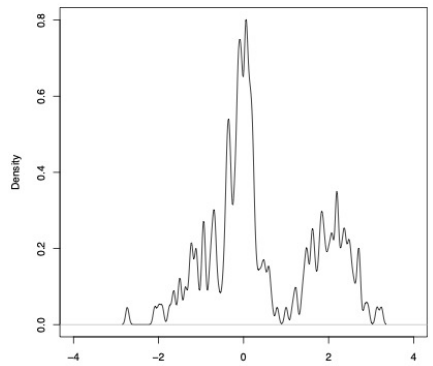
(a) Yes (b) No (c) Impossible to tell



(b) (2 pt) Are the two Gaussian kernel density plots below displaying exactly the same data? Circle only one answer.

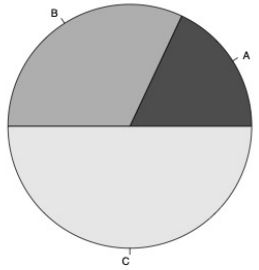
(a) Yes (b) No (c) Impossible to tell





(c) (2 pt) Which of the following can be determined from looking at this pie chart? Circle only one answer.

(a) Category B is twice as frequent as category A
(b) Category A is half as frequent as category B
(c) Category B is more frequent than category A
(d) All of the above
(e) None of the above

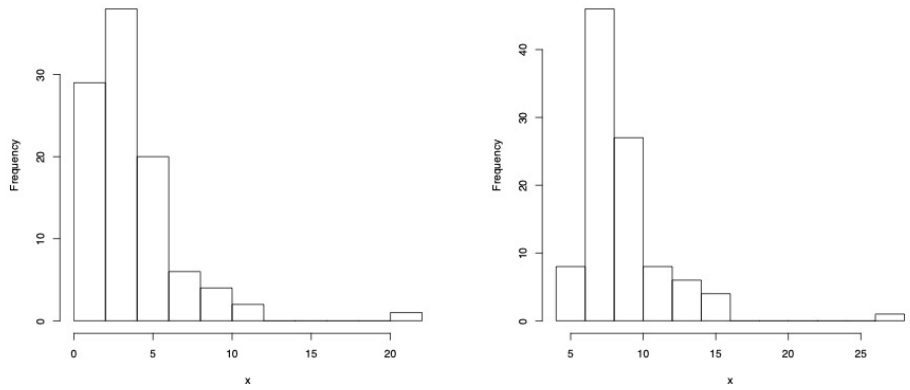


(e) (2 pt) You are given six lists, each of a few thousand numbers taking on values in the real line. Which of the following is the most effective way to visually compare the center and spread of the corresponding six distributions? Choose only one answer.

(a) Side-by-side box plots
(b) Side-by-side histograms

(a) (2 pt) Are the two histograms below displaying exactly the same data? Circle only one answer.

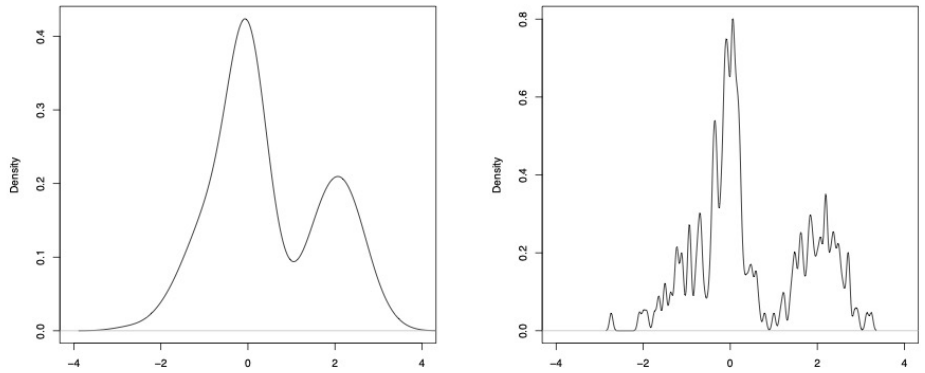
(a) Yes (b) No (c) Impossible to tell



There is a location shift between the two distributions.

(b) (2 pt) Are the two Gaussian kernel density plots below displaying exactly the same data? Circle only one answer.

(a) Yes (b) No (c) Impossible to tell

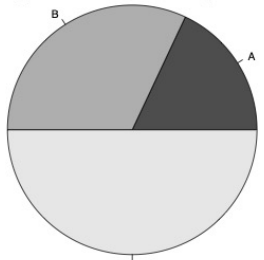


The density plots could be displaying the same data and look very different because of the bandwith, however, one cannot tell for sure.

(c) (2 pt) Which of the following can be determined from looking at this pie chart? Circle only one answer.

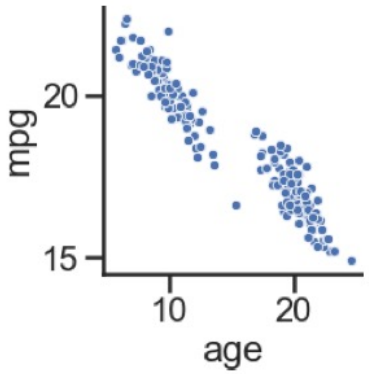
(a) Category B is twice as frequent as category A
(b) Category A is half as frequent as category B
(c) Category B is more frequent than category A
(d) All of the above
(e) None of the above

With angles/areas, it is very hard to precisely compare frequencies.



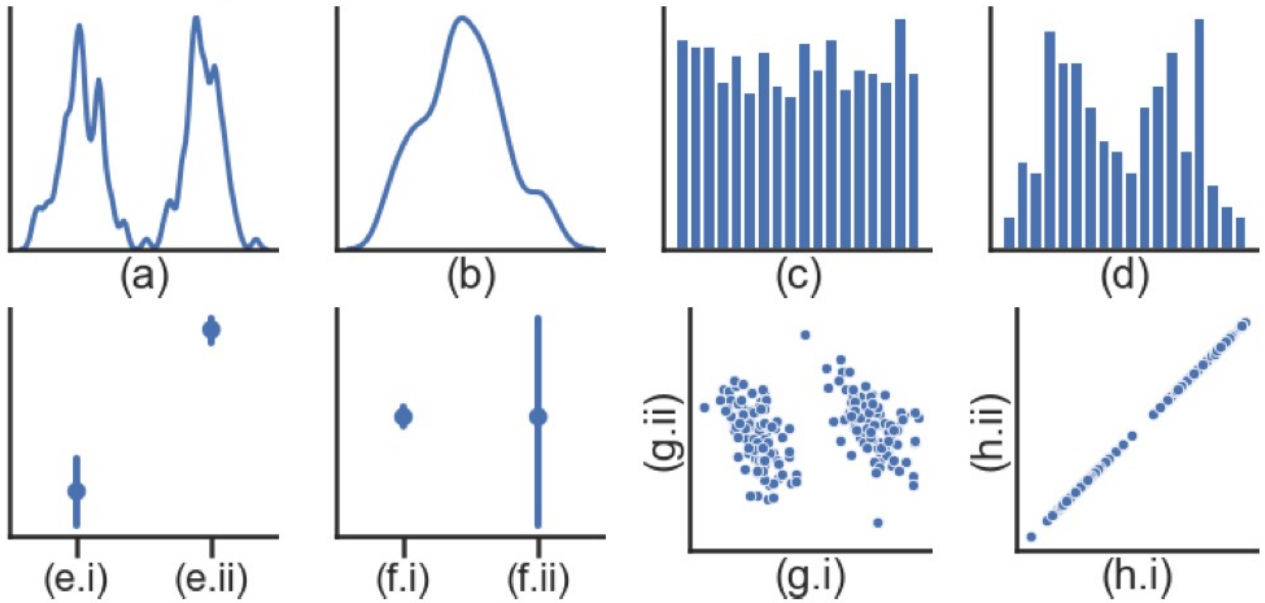
(e) (2 pt) You are given six lists, each of a few thousand numbers taking on values in the real line. Which of the following is the most effective way to visually compare the center and spread of the corresponding six distributions? Choose only one answer.

(a) Side-by-side box plots
(b) Side-by-side histograms



During a data analysis on car attributes, Sam created several plots. However, he has lost the axis labels for all of his plots except for the scatter plot shown on the left. Determine whether the plots below were generated from the same data. If so, mark the axis label that makes each plot consistent with the data in the scatter plot.

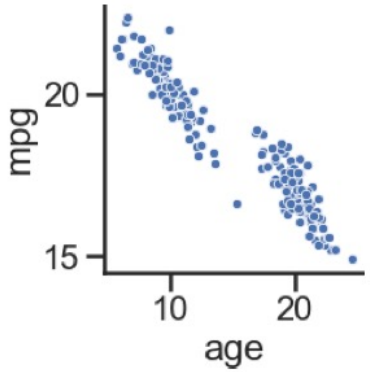
Assume that: The KDE plots use the same bandwidth, the histograms use the same number of bins, and point plots show the means of two columns and 95% confidence intervals. The axis limits for each plot were automatically chosen to display all plotted marks.



(a) (7 pt) Fill the missing axis labels of the 8 plots above using either `age` or `mpg` to make the plots consistent with the labeled scatter plot. For example, the first plot shows the distribution of `age`, so (a) should be filled in with `age`. If the plot cannot be generated from the data in either `age` or `mpg`, select Neither.

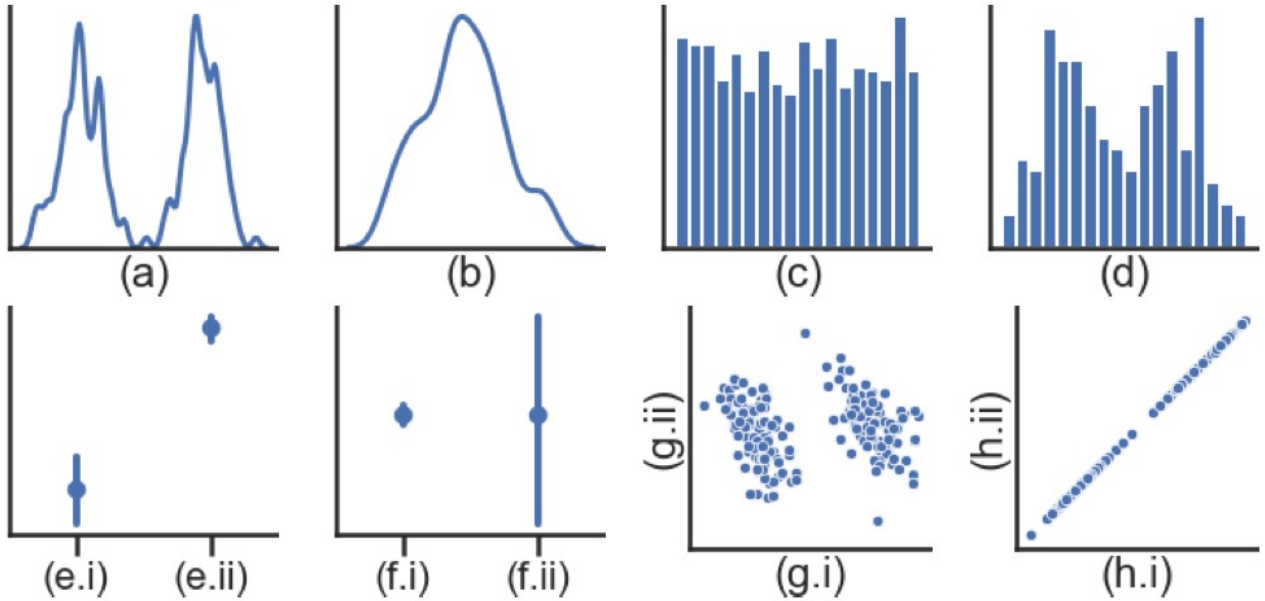
	(a)	(b)	(c)	(d)	(e.i)	(e.ii)	(f.i)	(f.ii)	(g.i)	(g.ii)	(h.i)	(h.ii)
age	●	○	○	○	○	○	○	○	○	○	○	○
mpg	○	○	○	○	○	○	○	○	○	○	○	○
Neither	○	○	○	○	○		○		○		○	

Note. A kernel density estimate (KDE) plot is **a method for visualizing the distribution of observations in a dataset**, analagous to a histogram.



During a data analysis on car attributes, Sam created several plots. However, he has lost the axis labels for all of his plots except for the scatter plot shown on the left. Determine whether the plots below were generated from the same data. If so, mark the axis label that makes each plot consistent with the data in the scatter plot.

Assume that: The KDE plots use the same bandwidth, the histograms use the same number of bins, and point plots show the means of two columns and 95% confidence intervals. The axis limits for each plot were automatically chosen to display all plotted marks.



	(a)	(b)	(c)	(d)	(e.i)	(e.ii)	(f.i)	(f.ii)	(g.i)	(g.ii)	(h.i)	(h.ii)
age	●	○	○	○	⊗	○	○	○	○	○	⊗	⊗
mpg	○	○	○	⊗	○	⊗	○	○	○	○	○	○
Neither	○	⊗	⊗	○	○		⊗		⊗		○	

(a-d): Notice that both age and mpg are bimodal, but the data have a gap in age and not mpg.
(e-g): The average age is lower than the average mpg.
(h): Either age for both axes or mpg for both axes is correct. The same variable plotted on both x and y axes will give a line with slope 1.

Note. A kernel density estimate (KDE) plot is **a method for visualizing the distribution of observations in a dataset**, analagous to a histogram.

A mysterious stranger on Sproul Plaza stops you on your way to class and claims that she has learned to flip any coin such that it lands on heads more often than the 50% you'd expect from random chance. To demonstrate, she takes a penny from her wallet, flips it 10 times, and gets heads nine times and only gets tails once.

(a) [4 Pts] The null hypothesis is that this was pure random chance, and that the probability of getting heads was 50% for each flip. What is the p-value under the null hypothesis of getting 1 or fewer tails out of 10 flips? You may leave your answer as a fraction if necessary.

(b) [4 Pts] Suppose the stranger flips the coin 28 more times, and they all end up heads. The resulting p value including all 38 flips under the null hypothesis is approximately $p_b = 10^{-10}$. Which of the following are true? Select all that apply.

- ☐ A. It is extremely unlikely that the stranger just happened to get 37 heads by randomly getting heads on 50/50 coin flips.
- ☐ B. p_b is the probability that the null hypothesis is true.
- ☐ C. $1 - p_b$ is the probability that the stranger has the skill to flip any arbitrary coin and get heads.
- ☐ D. If you flipped a fair coin 38 times, p_b is the chance that you'd get at least 37 heads by random chance.
- ☐ E. The stranger has proven beyond any reasonable doubt that she has the skill to flip any coin to land on heads with high probability.
- ☐ F. None of the above.

A mysterious stranger on Sproul Plaza stops you on your way to class and claims that she has learned to flip any coin such that it lands on heads more often than the 50% you'd expect from random chance. To demonstrate, she takes a penny from her wallet, flips it 10 times, and gets heads nine times and only gets tails once.

(a) [4 Pts] The null hypothesis is that this was pure random chance, and that the probability of getting heads was 50% for each flip. What is the p-value under the null hypothesis of getting 1 or fewer tails out of 10 flips? You may leave your answer as a fraction if necessary.

Solution: $\frac{11}{2^{10}}$
There is 1 way to get all heads, and n ways to get one tails. This is out of a total of 2^n permutations possible. Therefore the chances of getting 1 or fewer heads is $\frac{11}{2^{10}}$.

(b) [4 Pts] Suppose the stranger flips the coin 28 more times, and they all end up heads. The resulting p value including all 38 flips under the null hypothesis is approximately $p_b = 10^{-10}$. Which of the following are true? Select all that apply.

- ☐ **A. It is extremely unlikely that the stranger just happened to get 37 heads by randomly getting heads on 50/50 coin flips.**
- ☐ B. p_b is the probability that the null hypothesis is true.
- ☐ C. $1 - p_b$ is the probability that the stranger has the skill to flip any arbitrary coin and get heads.
- ☐ **D. If you flipped a fair coin 38 times, p_b is the chance that you'd get at least 37 heads by random chance.**
- ☐ E. The stranger has proven beyond any reasonable doubt that she has the skill to flip any coin to land on heads with high probability.
- ☐ F. None of the above.

Jane is a researcher in political science who is interested in understanding the effects of obtaining American citizenship on measures of patriotism and nationalism among American immigrants. She believes that granting citizenship to immigrants makes them more patriotic than similar immigrants who have not been granted citizenship yet.

To test her hypothesis, she filed a FOIA (Freedom of Information Act) request and gathered the following data on 1000 immigrants that applied for citizenship in 2015:

- Age.
- Sex.
- Country of origin.
- Last name.
- Year first arrived in the US.
- Occupation.
- Marital status.
- Number of children.
- Answer to a question about whether immigrant identifies as American or as someone from their country of origin.
- Answer to a question about where the immigrant traveled in the past year.
- A score, produced by Immigration and Customs Enforcement, used to determine eligibility for citizenship for each immigrant. Range: 1-100, Cutoff: 50.
- Favorite color.
- Date of birth.
- Race.

1. What hypothesis is Jane testing here?

2. What are the independent and dependent variables that Jane should identify in this study?

3. Jane wants to make causal inferences about how citizenship affects patriotism. Can she do this using the data that she has collected?

4. If the answer to **3** is "yes", then what kind of study, observational, quasi-experimental, or experimental should she use to answer the questions that she's interested in (be specific!). If the answer to **3** is "no", please explain why. Remember she can only use the data that she has collected, not hypothetical data.

5. Using the variables described above, explain how Jane should design her study. For example, one way to study a question related to favorite color and country of origin might be to designate favorite color as a dependent variable, country of origin as an independent variable and use age, sex, occupation and marital status as controls.