Practice Final
DSC 40A
Winter 2019

Exam is Saturday, March 16, 3-6pm

1. **Descriptive Statistics**

   Consider a data set satisfying **all** of the following requirements:

   - at least 1/4 of the data values are at least 8
   - at least 1/2 of the data values are at least 4
   - all data values are at least 2

   For each statistic given below, give an example of a data set that minimizes that statistic, among all data sets that satisfy the above requirements. Argue why the statistic cannot be made any smaller while satisfying all of the requirements for the data set.

   (a) mean

   (b) median

   (c) mode

   (d) midrange

2. **Descriptive Statistics**

   (a) Give an example of a data set for which removing one element changes the median but does not change the set of modes. Specify the data set, the element to remove, the old median, the new median, and the set of modes (same for old and new).

   (b) Give an example of a data set for which removing one element changes the set of modes but does not change the median. Specify the data set, the element to remove, the old set of modes, the new set of modes, and the median (same for old and new).

3. **Loss Functions**

For any data set with $d_1 \leq d_2 \leq \cdots \leq d_n$ and n even, define the first quartile Q_1 to be the median of the first $\frac{n}{2}$ data values, and the third quartile Q_3 to be the median of the last $\frac{n}{2}$ data values. Similarly, when n is odd, define the first quartile Q_1 to be the median of the first $\frac{n-1}{2}$ data values, and the third quartile Q_3 to be the median of the last $\frac{n-1}{2}$ data values. The *interquartile range* is a measure of spread defined as $Q_3 - Q_1$.

Find a loss function $L(h)$, defined in terms of Q_1 and Q_3, with the property that the minimum value of $L(h)$ equals the interquartile range. At what value of h is $L(h)$ minimized?

4. **Mean/Median Absolute Deviation from the Mean/Median**

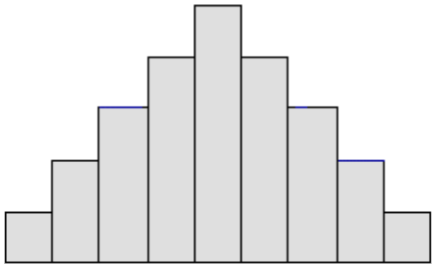
In class, we defined the *mean absolute deviation from the median* as a measure of the spread of a data set. This measure takes the absolute deviations, or differences, of each value in the data set from the median, and computes the mean of these absolute deviations. We can think of this one measure of spread as a member of a family of analogously defined measures of spread:

- mean absolute deviation from the median
- median absolute deviation from the median
- mean absolute deviation from the mean
- median absolute deviation from the mean

While all four of these measures capture the notion of spread, they do so in different ways, and so they may have different values for the same data set.

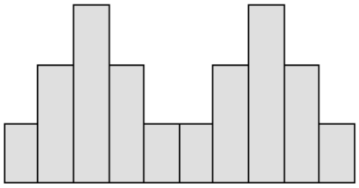
(a) For the data set whose histogram is shown below, draw a histogram showing the rough shape of the distribution of the absolute deviations from the mean. Which of these two measures is greater, or are they about the same?

- mean absolute deviation from the mean
- median absolute deviation from the mean



(b) For the data set whose histogram is shown below, draw a histogram showing the rough shape of the distribution of the absolute deviations from the median. Which of these two measures is greater, or are they about the same?

- mean absolute deviation from the median
- median absolute deviation from the median



5. **Big O, Theta, and Omega**

For each part, answer True or False, and justify your answer. All logarithms are base 2.

(a) $\sqrt{n^3} \in O(n^2)$.

(b) $2 \cdot 8^{\log(n^2)} \in \Theta(n^6)$.

(c) $\log(n) \in \Omega(\log(\log(n)))$.

(d) If f, g, and h are functions from the natural numbers to the non-negative real numbers with $f(n) \geq g(n)$ for all $n \geq 1$, $f(n) \in \Theta(h(n))$, and $g(n) \in \Theta(h(n))$, then $(f - g)(n)) \in \Theta(h(n))$.

(e) If f, g, and h are functions from the natural numbers to the non-negative real numbers with $f(n) \in \Theta(h(n))$ and $g(n) \in \Theta(h(n))$, then $(f * g)(n)) \in \Theta((h(n))^2)$.

6. **Regression**

Suppose that we generate data according to the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where each of the error terms ϵ_i has mean 0 and variance σ^2.

If we quadruple ($\times 4$) the variance of each error term ϵ_i, what effect does this have on...

(a) the mean value of y_i associated with a given x_i?

(b) the variance of the value of y_i associated with a given x_i?

(c) the value of β_0?

(d) the value of β_1?

7. **Regression**

Suppose we have a data set

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

to which we fit a line using linear regression. Is the line fitted to the original data set the same as the line fitted to the data set

$$(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}), (\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}), (x_3, y_3), \ldots, (x_n, y_n)?$$

If yes, provide a proof. If no, provide a counterexample.

8. **Regression**

A real estate agent wants to determine how to price a house, based on the square footage s, number of bedrooms t, and the average income of residents in the neighborhood u, in thousands of dollars. A model for the total price of the house P, in thousands of dollars, based on combinations of these variables is

$$P = c_0 + c_1 s + c_2 t + c_3 u + c_4 s^2 + c_5 t^2 + c_6 u^2 + c_7 st + c_8 su + c_9 tu.$$

The goal is to approximate the values of the constants c_i based on data from houses that have sold recently. The real estate agent collects the following data:

- A 3-bedroom house with 1800 square feet, in a neighborhood where the average income was $\$70,000$ sold for $\$800,000$.
- A 4-bedroom houe with 3000 square feet, in a neighborhood where the average income was $\$95,000$ sold for $\$1,100,000$.
- A 3-bedroom house with 2200 square feet, in a neighborhood where the average income was $\$50,000$ sold for $\$625,000$

Specify the design matrix X, observation vector $\vec{y}$, and the form of the parameter vector $\vec{b}$ that correspond to this scenario. You do not need to simplify or perform any calculations.

9. **Regression**

Let $\vec{b}$ be the specific parameter vector that satisfies $X^T X\vec{b} = X^T\vec{y}$. Consider the following three sums of squares:

$$||X\vec{b}||^2 \qquad ||\vec{y} - X\vec{b}||^2 \qquad ||\vec{y}||^2$$

These three quantities are related in a simple way that is of importance in statistics. What is this relationship? Prove why the relationship is true in general.

10. **Counting**

Give an exact answer for each part. You can leave your answer as an unsimplified expression in terms of combinations, permutations, factorials, etc.

Note: A digit is defined as an integer 0 through 9, and strings of digits can start with a 0.

(a) How many different initials can a person have if they have a first name, a middle name, and a last name?

(b) How many length 6 bitstrings with an equal number of zeroes and ones start with a 1 and end with a 1?

(c) How many ways are there to arrange 3 women (Ada, Grace, Eva) and 3 men (Alan, Ron, Sergey) in a line so that a woman is first and a woman is last?

(d) How many different words can be made by rearranging the letters in the word LETTERS?

(e) How many sandwiches can you order from a restaurant, if sandwiches must include one bread (wheat, white, italian, or rye), one meat (turkey, roast beef, ham, or tuna) and one cheese (swiss, cheddar, or provolone), plus any number of toppings (lettuce, tomato, onion, pickles, mustard, mayonnaise, pesto)?

(f) You order a sandwich with wheat bread, turkey, provolone, lettuce, tomato, and pesto. Your sandwich will have bread on the bottom and bread on the top, but if the other ingredients can be piled on in any order, how many ways are there to make your sandwich?

(g) A California license plate follows the format $DLLLDDD$, where D represents a digit and L represents an upper case letter. For example, a valid California license plate is $6AMB205$. How many California license plates are possible?

(h) A company requires its employees to create account passwords that are exactly six characters in length. Passwords can contain upper case letters, lower case letters, and numbers. How many passwords are possible in this system?

(i) How many strings of length ten consist of two different digits alternating? For example, 4747474747.

(j) How many strings of length ten consist of two copies of each odd digit? For example, 7539715139.

11. **Counting**

In each of the following problems, a hand of five cards will be dealt from a standard deck of cards, with thirteen cards in each of four suits, and no jokers or wild cards. A hand is a *set* of five cards, so the order in which the cards are dealt does not

matter. Say how many different hands of the following types are possible. You do not need to simplify your answer at all.

(a) Straight: A hand where the numbers of the cards are five consecutive integers (with Jack = 11, Queen = 12, King = 13, and Ace counting as 1 or 14).

(b) Three of a kind: A hand with three cards of one number, one card of a second number, and one card of a third number.

(c) Two pair: A hand with two cards of one number, two cards of a second number, and one card of a third number.

12. **Probability**

(a) Suppose there are two bowls of cookies. Bowl 1 has 10 chocolate chip and 30 peanut butter cookies, while Bowl 2 has 20 of each. You pick a bowl at random, and then pick a cookie at random. You get a peanut butter cookie. How likely is it that you picked out Bowl 1?

(b) Suppose there are two bowls of cookies. Bowl 1 has 10 chocolate chip and 30 peanut butter cookies, while Bowl 2 has 80 of each. You pick a bowl at random, and then pick a cookie at random. You get a peanut butter cookie. How likely is it that you picked out Bowl 1? Explain how your answer compares to the previous question and why.

13. **Probability**

A child walks into a library and randomly rearranges all n books on a bookshelf.

(a) What is the expected number of books that wind up in their correct place on the bookshelf? Show your work.

(b) What is the expected number of pairs of books that wind up in the correct relative order? Show your work.

(c) What is the expected number of books that wind up somewhere to the left of their original position? Show your work.

14. **Probability**

Consider the sample space $S = \{a, b, c, d, e\}$ with the uniform probability distribution. Define nontrivial events A, B, and C such that A and B are conditionally independent given C. A nontrivial event is one for which the probability is strictly between 0 and 1, without equalling 0 or 1.

15. **Probability**

In a dice game, players take turns rolling and adding to their total score. A player's score in each round is determined by rolling a 6-sided die and a 4-sided die, subtracting the two resulting numbers, and squaring the difference. Player A went first and has already finished the last round, and player B has one more turn remaining. At this point, player A has 43 points, and player B has 40 points. Who do you think will win the game in the end? What score to you expect player B to have when the game is over?

Give values of a and b such that before player B's last turn,

- player A has a points,
- player B has b points,
- the expected final score of player A is less than the expected final score of player B, and
- the probability that player A wins is greater than the probability that player B wins.

16. **Sampling**

(a) A population consists of n people. A sample of k people is drawn at random **with replacement** from the population. What is the expected number of people in the sample who appear more than once?

(b) A population consists of n people. A sample of k people is drawn at random **without replacement** from the population. What is the probability that individuals were chosen in order of increasing height? Assume no two individuals in the population have the same height.

(c) A population consists of n people. A sample of k people is drawn at random **with replacement** from the population. A second sample of k people is drawn at random **without replacment** from the population. What is the probability that the two samples are exactly the same (they contain the same individuals in the same order)?

17. **k-Means Clustering**

We are given the following data and want to find $k = 3$ clusters.

$$x^{(1)} = (2, 10), \quad x^{(2)} = (2, 5), \quad x^{(3)} = (8, 4), \quad x^{(4)} = (5, 8)$$
$$x^{(5)} = (7, 5), \quad x^{(6)} = (6, 4), \quad x^{(7)} = (1, 2), \quad x^{(8)} = (4, 9)$$

Suppose we randomly select $x^{(1)}$, $x^{(4)}$ and $x^{(7)}$ as cluster centers. Trace through one iteration of Lloyd's algorithm and find the new cluster centers after this first iteration. Show that the cost function has decreased with this iteration.