Practice Final Solutions
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

   Solution: The data set 2, 2, 2, 2, 4, 4, 8, 8 makes each value as small as possible subject to the requirements. Since each data value is as small as possible, the mean is as small possible. Note that only the relative proportion of each value matters: we could have used a data set of 2, 2, 4, 8 and gotten the same answer.

   (b) median

   Solution: The data set 2, 2, 2, 2, 4, 4, 8, 8 makes each value as small as possible subject to the requirements. Since each data value is as small as possible, in particular the middle two numbers are as small as possible, so the median is as small as possible.

(c) mode

Solution: The data set 2, 2, 2, 2, 4, 4, 8, 8 has a mode of 2. Since each value in the data set must be at least 2, it would be impossible to have a mode smaller than 2.

(d) midrange

Solution: The data set 2, 2, 2, 2, 4, 4, 8, 8 makes each value as small as possible subject to the requirements. Since each data value is as small as possible, in particular, the smallest and largest numbers (which determine the midrange) are as small as possible, so the midrange is as small as possible.

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).

Solution: Start with the data set 1, 1, 1, 2, 3, 4, which has mode 1 and median 1.5. When we remove the first element, the data set becomes 1, 1, 2, 3, 4, which has mode 1 and and median 2. The median has changed but the mode has stayed the same.

(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).

Solution: Start with the data set 1, 1, 3, 3, 5, which has modes 1 and 3 and median 3. When we remove the first element, the data set becomes 1, 3, 3, 5, which has mode 3 and and median 3. The set of modes has changed but the median has stayed the same.

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?

Solution: We know that the loss function $L(h) = \max(|h - d_1|, |h - d_n|)$ is minimized at $h = \frac{d_1+d_n}{2}$ and has a minimum value of $L\left(\frac{d_1+d_n}{2}\right) = \frac{d_n-d_1}{2}$.

Substituting Q_1 for d_1 and Q_3 for d_n gives that the loss function $L(h) = \max(|h - Q_1|, |h - Q_3|)$ is minimized at $h = \frac{Q_1+Q_3}{2}$ and has a minimum value of $L\left(\frac{Q_1+Q_3}{2}\right) = \frac{Q_3-Q_1}{2}$.

This is almost what we want, except we want the minimum value to be Q_3-Q_1, which we can achieve by multiplying the loss function by two. Therefore, the loss function $L(h) = 2\max(|h - Q_1|, |h - Q_3|)$ has the property that the minimum value of $L(h)$ equals the interquartile range. This $L(h)$ is minimized at the midpoint of the interquartile range, $h = \frac{Q_1+Q_3}{2}$.

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

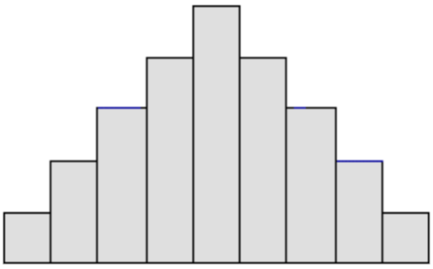
In class, we defined the *mean absolute deviation from the median* as a measure of the spread of a data set. This measure takes the absolute deviations, or differences, of each value in the data set from the median, and computes the mean of these absolute deviations. We can think of this one measure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

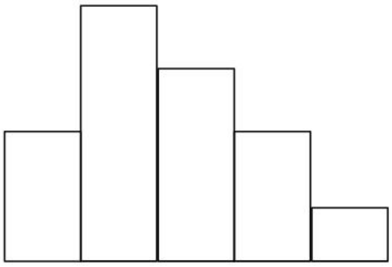
While all four of these measures capture the notion of spread, they do so in different ways, and so they may have different values for the same data set.

(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



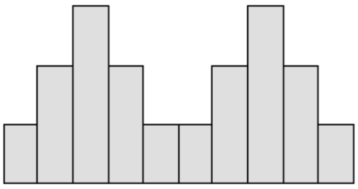
Solution: The mean is in the center of the histogram since it is roughly symmetric. The histogram of absolute deviations from the mean is shaped approximately like this:



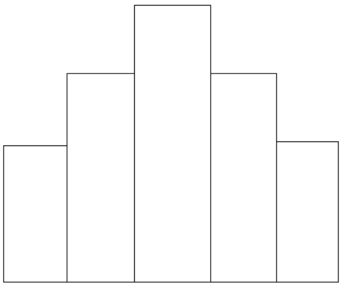
Since the histogram of absolute deviations from the mean is right skewed, the mean of this distribution is greater than the median. Therefore, the mean absolute deviation from the mean should be greater than the median absolute deviation from the mean.

(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



Solution: The median is in the center of the histogram since it is roughly symmetric. The histogram of absolute deviations from the median is



Since the histogram of absolute deviations from the median is roughly symmetric, we know its mean value is about the same as its median. Therefore, the mean absolute deviation from the median should be about the same as the 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) $\sqrt{n^3} \in O(n^2)$.

Solution: True. As $n \to \infty$, the limit of $\frac{\sqrt{n^3}}{n^2}$ is 0, which says $\sqrt{n^3} \in O(n^2)$.

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

Solution: True. $2 \cdot 8^{\log(n^2)} = 2 \cdot \left(2^3\right)^{\log(n^2)} = 2 \cdot 2^{3\log(n^2)} = 2 \cdot 2^{\log\left((n^2)^3\right)} = 2 \cdot 2^{\log(n^6)} = 2 \cdot n^6$. This is just a constant multiple of n^6 so both functions are in the same Θ class.

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

Solution: True. We can satisfy the definition of Ω with the inequality $\log(n) \geq 1 \cdot \log(\log(n))$ for all $n > 2$.

(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 $(f - g)(n)) \in \Theta(h(n))$.

Solution: False. Here is a counterexample: $f(n) = 3n + 1, g(n) = 3n, h(n) = n$.

(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 $(f * g)(n)) \in \Theta((h(n))^2)$.

Solution: True. Since $f(n) \in \Theta(h(n))$ and $g(n) \in \Theta(h(n))$, that means we can find constants $C_1, D_1, k_1, C_2, D_2, k_2$ so that

$$D_1 \cdot h(n) \leq f(n) \leq C_1 \cdot h(n) \text{ for all } n > k_1 \text{ and}$$
$$D_2 \cdot h(n) \leq g(n) \leq C_2 \cdot h(n) \text{ for all } n > k_2.$$

If we multiply these inequalities, this means that for all $n > max(k_1, k_2)$,

$$D_1 D_2 \cdot (h(n))^2 \leq (f * g)(n) \leq C_1 C_2 \cdot (h(n))^2,$$

which says $(f * g)(n)) \in \Theta((h(n))^2)$.

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 (×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?

Solution: No effect. Changing the spread of the values of y_i associated with x_i does not affect the mean.

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

Solution: Quadruple. For a fixed x_i, the variance of the y_i values is the same as the variance of the error terms ϵ_i because $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and $\beta_0 + \beta_1 x_i$ is constant.

(c) the value of β_0?

Solution: No effect. β_0 is the intercept of the true line. Increasing the spread of points around the line does not change where the true line falls.

(d) the value of β_1?

Solution: No effect. β_1 is the slope of the true line. Increasing the spread of points around the line does not change where the true line falls.

7. **Regression**

Suppose we have a data set

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

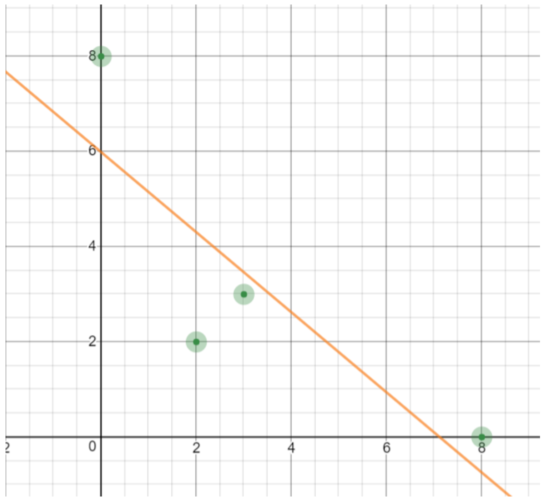
to which we fit a line using linear regression. Is the line fitted to the original data set the same as the line fitted to the data set

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

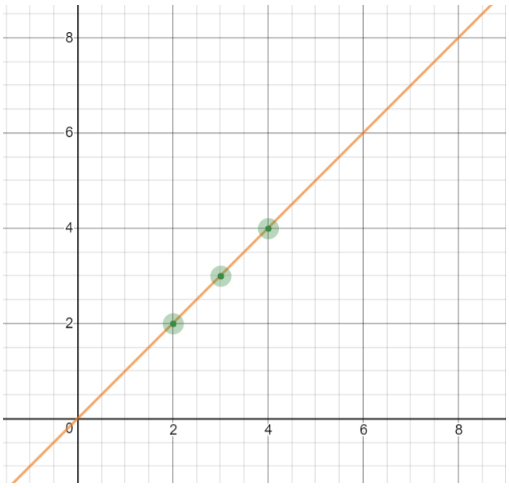
If yes, provide a proof. If no, provide a counterexample.

Solution:

No, the lines are not the same in general. A counterexample is the data set $(0, 8), (8, 0), (2, 2), (3, 3)$, pictured below.



When we transform the data set by averaging the first two points, the data set becomes $(4,4),(4,4),(2,2),(3,3)$, pictured below.



This shows that the line fitted to the original data set may be different than the line fitted to the transformed data set.

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.

Solution: This corresponds to

$$X = \begin{bmatrix} 1 & 1800 & 3 & 70 & (1800)^2 & (3)^2 & (70)^2 & 1800*3 & 1800*70 & 3*70 \\ 1 & 3000 & 4 & 95 & (3000)^2 & (4)^2 & (95)^2 & 3000*4 & 3000*95 & 4*95 \\ 1 & 2200 & 3 & 50 & (2200)^2 & (3)^2 & (50)^2 & 2200*3 & 2200*50 & 3*50 \end{bmatrix},$$

$$\vec{b} = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \\ c_9 \end{bmatrix}, \vec{y} = \begin{bmatrix} 800 \\ 1,100 \\ 625 \end{bmatrix}.$$

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.

Solution: The relationship is

$$||b||^2 = ||b - A\widehat{x}||^2 + ||A\widehat{x}||^2.$$

We have $A\widehat{x} = \widehat{b}$, which is the orthogonal projection of b onto $Col(A)$. By the Orthogonal Decomposition Theorem, $b = \widehat{b} + (b - \widehat{b})$ where $\widehat{b}$ is in $Col(A)$ and $(b - \widehat{b}) \perp Col(A)$. Since $(b - \widehat{b}) \perp Col(A)$ and $\widehat{b}$ is one vector in $Col(A)$, we know $(b - \widehat{b}) \perp \widehat{b}$. Replacing $\widehat{b}$ with $A\widehat{x}$, this says $(b - A\widehat{x}) \perp A\widehat{x}$. Then applying the Pythagorean Theorem to these two vectors gives

$$||(b - A\widehat{x}) + A\widehat{x}||^2 = ||b - A\widehat{x}||^2 + ||A\widehat{x}||^2,$$

and simplifying the left hand side proves the claimed relationship

$$||b||^2 = ||b - A\widehat{x}||^2 + ||A\widehat{x}||^2.$$

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?

Solution: 26^3, since each initial can be one of 26 letters and initials are an ordered sequence of 3 letters.

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

Solution: $\binom{4}{1} = 4$, since we just need to choose one of the four middle positions to have a 1, which determines the string.

(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?

Solution: There are 3 women who could be first, then any of the remaining 2 can be last. The other 4 people can be arranged in any order in the middle positions. This gives $3 * 4! * 2 = 144$. Notice that this is a similar question to the previous question, except people are distinguishable and bits are not, so the answer is different.

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

Solution: First choose two of the seven positions for the two E's, then choose two of the remaining positions for the two T's, then choose one of the remaining positions for the L, then choose one of the remaining positions for the R, then choose one of the remaining positions for the S. This gives $\binom{7}{2}\binom{5}{2}\binom{3}{1}\binom{2}{1}\binom{1}{1} = 1260$.

(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)?

Solution: There are 4 bread options, 4 meat options, 3 cheese options, 2 lettuce options (yes/no), 2 tomato options (yes/no), 2 onion options (yes/no), 2 pickle options (yes/no), 2 mustartd options (yes/no), 2 mayonnaise options (yes/no), and 2 pesto options (yes/no). This gives $4 * 4 * 3 * 2^7 = 6144$.

(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?

Solution: Aside from the bread, there will be 5 ingredients in your sandwich. The number of ways to make the sandwich is the same as the number of ways to reorder or permute these 5 ingredients. The number of permutations of 5 things is $5! = 120$.

(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?

Solution: There are 26 upper case letters and 10 digits. Using the number of options in each position gives $10 * 26 * 26 * 26 * 10 * 10 * 10 = 175760000$.

(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?

Solution: There are 26 upper case letters, 26 lower case letters, and 10 digits, so each character has $26 + 26 + 10 = 62$ options. The number of sequences of length 6 is therefore $62^6 = 56800235584$.

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

Solution: We have 10 options for the first digit of the string, and 9 options for the second digit of the string, since the two digits must be different. These two choices determine the rest of the string, so the number of possible strings is $10 * 9 = 90$.

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

Solution: Of the ten positions, we must choose two of them to have 1's, then two of the remaining 8 positions to have 3's, then two of the remaining 6 positions to have 5's, then two of the remaining 4 positions to have 7's. At this point, there are only two remaining positions, which must necessarily have 9's. The number of ways to choose these positions is $\binom{10}{2} * \binom{8}{2} * \binom{6}{2} * \binom{4}{2} = 113400$.

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).

Solution: We could start the straight at any number between 1 and 10. This means there are 10 possible sets of five numbers which could make up the straight. For each of the five numbers in the straight, we must use exactly one of that number and there are four suits to pick from. Thus, the total number of straights is $10 * 4^5 = 10,240$.

(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.

Solution: Of the thirteen kinds of numbers, we must pick one to have three cards and two of the remaining to have one card. That is $13 * \binom{12}{2}$ possibilities. Then, to choose the suits, we must pick three of the four cards with the first number and one of four cards for each of the other two numbers. That is $\binom{4}{3}\binom{4}{1}\binom{4}{1}$. So the total is $13 * \binom{12}{2}\binom{4}{3}\binom{4}{1}\binom{4}{1} = 54,912$ possible hands.

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

Solution: We must pick two numbers to have two of, and one of the remaining numbers to have one of, for $\binom{13}{2} * 11$ possible combinations of numbers. Then to pick suits for each of the pairs, we pick two of the four cards with that number, and to pick a suit for the remaining card, we pick one of the four cards with that number. This gives $\binom{4}{2}\binom{4}{2}\binom{4}{1} = 6 * 6 * 4$ possibile ways to choose the suits. So the total number of hands is $(\binom{13}{2} * 11) * 6 * 6 * 4 = 123,552$.

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?

Solution:

There are two ways to do this problem. One solution uses a sample space of all 80 cookies in both bowls. Each cookie is equally likely to be drawn so there is a uniform distribution on this sample space.

$$\begin{aligned} P(\text{bowl 1}|\text{pb}) &= \frac{P(\text{bowl 1 and pb})}{P(\text{pb})} \\ &= \frac{30/80}{50/80} \\ &= \frac{3}{5} \end{aligned}$$

Another solution uses Bayes Theorem.

$$\begin{aligned}
P(\text{bowl }1|\text{pb}) &= \frac{P(\text{pb}|\text{bowl }1)P(\text{bowl }1)}{P(\text{pb}|\text{bowl }1)P(\text{bowl }1)+P(\text{pb}|\text{bowl }2)P(\text{bowl }2)} \\
&= \frac{\frac{30}{40}*\frac{1}{2}}{\frac{30}{40}*\frac{1}{2}+\frac{20}{40}*\frac{1}{2}} \\
&= \frac{3/4}{5/4} \\
&= \frac{3}{5}
\end{aligned}$$

(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.

Solution: For this problem, only the Bayes Theorem approach above works, since we do not have a uniform distribution on the sample space of all cookies. Certain cookies (those in bowl 1) are more likely to be picked than others.

The solution using Bayes Theorem is

$$\begin{aligned}
P(\text{bowl }1|\text{pb}) &= \frac{P(\text{pb}|\text{bowl }1)P(\text{bowl }1)}{P(\text{pb}|\text{bowl }1)P(\text{bowl }1)+P(\text{pb}|\text{bowl }2)P(\text{bowl }2)} \\
&= \frac{\frac{30}{40}*\frac{1}{2}}{\frac{30}{40}*\frac{1}{2}+\frac{20}{40}*\frac{1}{2}} \\
&= \frac{3/4}{5/4} \\
&= \frac{3}{5}
\end{aligned}$$

Notice that this is the same as the previous question because in both question, the bowls have the same distribution of cookie types. The actual number of cookies is the only thing that changes, but the probabilities used in Bayes Theorem are all the same.

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.

Solution: Let X be the number of books that wind up in the correct position. Define a random variable X_i for each book $i = 1, 2, \ldots, n$. X_i should be 1 if book i is placed in its proper position, and 0 if not. Then $E(X_i) = 1 * P(X_i = 1) + 0 * P(X_i = 0) = P(X_i = 1)$, which is the probability that book i ends up in its correct place. This probability is $\frac{1}{n}$ because only one of the n possible positions is correct, and all positions are equally likely.

Therefore, $X = \sum_{i=1}^{n} X_i$, and using linearity of expectation,

$$E(X) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{1}{n} = n * \frac{1}{n} = 1.$$

We expect exactly one book to end up in its correct place, regardless of the nuber of books on the bookshelf.

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

Solution: Let X be the number of pairs of books that wind up in the correct relative order. Suppose the books are numbered $1, 2, \ldots, n$ from left to right, before they are rearranged. Define a random variable $X_{i,j}$ for each pair of books i and j with $i < j$. $X_{i,j}$ should be 1 if book i is placed to the left of book j (the correct relative order), and 0 if book i is placed to the right of book j. Then $E(X_{i,j}) = 1 * P(X_{i,j} = 1) + 0 * P(X_{i,j} = 0) = P(X_{i,j} = 1)$, which is the probability that book i ends up to the left of book j. This probability is $\frac{1}{2}$ because the books are rearranged randomly, so it is equally likely that i comes before j as it is that j comes before i.

There are $\binom{n}{2}$ many ways to select a pair of books i and j with $i < j$, so using linearity of expectation,

$$E(X) = \sum_{i,j:i<j} E(X_{i,j}) = \sum_{i,j:i<j} \frac{1}{2} = \frac{1}{2} * \binom{n}{2}.$$

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

Solution: Let X be the number of books that wind up somewhere to the left of their original position. Suppose the books are numbered $1, 2, \ldots, n$ from left to right, before they are rearranged. Define a random variable X_i for each book. X_i should be 1 if book i is placed to left of position i, and 0 otherwise. Then $E(X_i) = 1 * P(X_i = 1) + 0 * P(X_i = 0) = P(X_i = 1)$, which is the probability that book i ends up to the left of position i. Since book i has an equal probability of being placed in any of the n positions, and there are $i - 1$ positions to the left of position i, the probability that book i ends up to the left of position i is $\frac{i-1}{n}$.

Therefore, $X = \displaystyle\sum_{i=1}^{n} X_i$, and using linearity of expectation,

$$E(X) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \frac{i-1}{n} = \frac{1}{n}\sum_{i=1}^{n}(i-1)$$

We can evaluate $\displaystyle\sum_{i=1}^{n}(i-1) = 0 + 1 + 2 + \cdots + (n-2) + (n-1)$ by pairing the first term (0) with the last term $(n-1)$, the second term (1) with the second to last term $(n-2)$ and so on. The total of each pair is $n-1$, and there are $\frac{n}{2}$ pairs, so the sum evaluates to $(n-1) * \frac{n}{2}$. Plugging this in to our formula for $E(X)$ gives

$$E(X) = \frac{1}{n}\sum_{i=1}^{n}(i-1) = \frac{1}{n} * (n-1) * \frac{n}{2} = \frac{n-1}{2}.$$

This makes sense when paired with our answer to part (a). Of the n books, we expect one book to wind up where it started, and for the other $n-1$ books, we expect half of them to end up to the left of their starting position and half of them to end up to the right of their starting position.

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.

Solution: Let $A = \{c, d, e\}, B = \{b, c\}, C = \{a, b, c, d\}$. Then

$$\begin{aligned}
P((A \cap B)|C) &= \frac{P(A \cap B \cap C)}{P(C)} \\
&= \frac{P(\{c\})}{P(\{a, b, c, d\})} \\
&= \frac{1/5}{4/5} \\
&= \frac{1}{4} \\
P(A|C) &= \frac{P(A \cap C)}{P(C)} \\
&= \frac{P(\{c, d\})}{P(\{a, b, c, d\})} \\
&= \frac{2/5}{4/5} \\
&= \frac{1}{2} \\
P(B|C) &= \frac{P(B \cap C)}{P(C)} \\
&= \frac{P(\{b, c\})}{P(\{a, b, c, d\})} \\
&= \frac{2/5}{4/5} \\
&= \frac{1}{2}
\end{aligned}$$

This says that $P((A \cap B)|C) = P(A|C) * P(B|C)$ so the definition of conditional independence is satisfied.

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.

Solution:

	1	2	3	4
1	0	1	4	9
2	1	0	1	4
3	4	1	0	1
4	9	4	1	0
5	16	9	4	1
6	25	16	9	4

According to the table of possible outcomes above (the sample space), the expected value of a single roll is

$$0 * \frac{4}{24} + 1 * \frac{7}{24} + 4 * \frac{6}{24} + 9 * \frac{4}{24} + 16 * \frac{2}{24} + 25 * \frac{1}{24} \approx 5.17.$$

Player B's expected score at the end of the game is $40 + 5.17 = 45.17$, which is more than player A's score of 40, so we expect that player B will win with a final score of 45.17. Additionally, the probability that player B scores at least 3 points on the last turn is $\frac{13}{24}$ according to the table above, so player B is more likely than not to win.

If we let $a = 43$ and $b = 38$, then the expected final score of player B would be $38 + 5.17 = 43.17$, which is more than player A's final score. However, the probability of player B winning is only $\frac{7}{24}$ since it is rare to get a score of at least 5.

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?

Solution: Let X be the number of people who appear more than once. Suppose each person has a number $i = 1, 2, \dots, n$. Define the random variable X_i to be 1 if person i was included in the sample 2 or more times, and 0 if person i was included in the sample 0 or 1 times. It is actually easier to figure out $P(X_i = 0)$ than $P(X_i = 1)$. The probability that person i was never in the sample is $\left(\frac{n-1}{n}\right)^k$ because we need to have picked any of the other $n-1$ people all k times. The probability that person i appeared in the sample exactly once is $C(k,1) * \frac{1}{n} * \left(\frac{n-1}{n}\right)^{k-1}$ because we need to pick which of the k positions will be person i, then we need to have person i in that position with probability $\frac{1}{n}$ and we need to have someone else in each of the other $k-1$ positions. This means $P(X_i = 0) = \left(\frac{n-1}{n}\right)^k + C(k,1) * \frac{1}{n} * \left(\frac{n-1}{n}\right)^{k-1}$, so

$$E(X_i) = P(X_i = 1) = 1 - \left(\left(\frac{n-1}{n}\right)^k + C(k,1) * \frac{1}{n} * \left(\frac{n-1}{n}\right)^{k-1}\right).$$

Therefore, $X = \sum_{i=1}^{n} X_i$, and using linearity of expectation,

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} E(X_i) \\ &= \sum_{i=1}^{n} 1 - \left(\left(\frac{n-1}{n}\right)^k + C(k,1) * \frac{1}{n} * \left(\frac{n-1}{n}\right)^{k-1}\right) \\ &= n * \left(1 - \left(\left(\frac{n-1}{n}\right)^k + C(k,1) * \frac{1}{n} * \left(\frac{n-1}{n}\right)^{k-1}\right)\right) \end{aligned}$$

(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.

Solution: We can use a sample space of the possible orderings of the k selected individuals. There are $k!$ orders in which the k individuals could have been chosen. Only one ordering has them in increasing height order. Since each ordering is equally likely, the probability is therefore $\frac{1}{k!}$. Notice that the value of n is not a factor in our answer: it doesn't matter how many people you are sampling from.

(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)?

Solution: Take the second sample, the one drawn without replacement, and consider the probability that the first sample matches it. This probability is $\left(\frac{1}{n}\right)^k$ since there is a $\frac{1}{n}$ chance of each of the k individuals being exactly the same as the corresponding individual in the other sample.

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.

Solution: First, we assign $x^{(1)}, x^{(4)}, x^{(7)}$ to be our k cluster centers. Let $\mu_1 = x^{(1)}, \mu_2 = x^{(4)}, \mu_3 = x^{(7)}$.

Then, we assign each $x^{(i)}$ to a cluster. Let C_j denote the set of points belonging to the cluster centered at μ_j. So, $x^{(i)}$ belongs to the set C_j if the closest cluster center to $x^{(i)}$ is μ_j. We get

$$\begin{aligned} C_1 &= \{x^{(1)}\} \\ C_2 &= \{x^{(4)}, x^{(3)}, x^{(5)}, x^{(6)}, x^{(8)}\} \\ C_3 &= \{x^{(7)}, x^{(2)}\}. \end{aligned}$$

Now, we calculate the cost for each cluster center. Recall that $\text{cost}(\mu_j)$ is the total squared distance of each $x^{(i)}$ in C_j to μ_j. Calculating these distances gives

$$\begin{aligned}\text{cost}(\mu_1) &= 0\\ \text{cost}(\mu_2) &= 57\\ \text{cost}(\mu_3) &= 10\,.\end{aligned}$$

Summing these values gives us $\text{cost}(\mu_1, \mu_2, \mu_3) = 0 + 57 + 10 = 67$. The next step is to move each μ_j to the average coordinate of points in C_j. So, we make the following changes:

$$\begin{aligned}\mu_1 &= \left(\frac{2}{1}, \frac{10}{1}\right) = (2, 10)\\ \mu_2 &= \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right) = (6, 6)\\ \mu_3 &= \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = \left(\frac{3}{2}, \frac{7}{2}\right)\,.\end{aligned}$$

Finally, we can compute the new clusters and see the new values for the cost function are

$$\begin{aligned}\text{cost}(\mu_1) &= 5\\ \text{cost}(\mu_2) &= 19\\ \text{cost}(\mu_3) &= 5\,.\end{aligned}$$

Therefore, $\text{cost}(\mu_1, \mu_2, \mu_3) = 5 + 19 + 5 = 29$. Indeed, the cost function has decreased with this iteration as $29 < 67$.