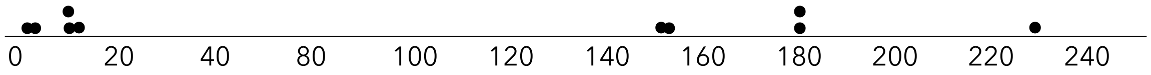
Assignment #12: Mean and median Name: ______________________________

Situation: On September 19, 2014 I went to amazon.com and searched for the phrase "probability and statistics." Sorted by relevance, the prices for the top 10 books under this search were (rounded to the nearest dollar and listed from cheapest to most expensive): **3, 4, 10, 10, 11, 152, 153, 180, 180, 227**

Here are those prices displayed as a dotplot:



1. Suppose we want to find (or calculate) a single number that best represents all ten book prices. We could decide to choose:

   a) 10 = the number of book prices in the dataset
   b) 3 (or 227)= the cheapest (or most expensive) book
   c) 224 = the difference in price from the cheapest to the most expensive book

   d) $115 = \frac{3+227}{2} =$ The midpoint between the minimum and maximum

   e) $81.5 = \frac{11+152}{2} =$ The midpoint between the middle two values = the median

   f) $93 = \frac{3+4+...+227}{10} =$ The average = the mean

   g) $86 = \frac{10+10+...+180}{10} =$ 20% trimmed mean (the avg. after we delete the lowest 2 and highest 2 values)

   h) $89.6 = \frac{10+10+10+10+11+152+153+180+180+180}{10} =$ The winsorized mean (the avg. after we replace the lowest two and highest two values with the next lowest and highest.

   i) 12-151 = Any of these numbers have an equal number of book prices to the left and right on the dotplot.

   j) Another number (using some method I didn't think of)

   Write down your choice for the number that best represents all ten book prices: ______________

2. How do we decide which number is best? Well, that entirely depends on our intentions (i.e., the purpose of the number we want to calculate).

   Suppose we want to find the number that best represents the center of our ten book prices. Obviously, values like 3, 12, 200, or 227 would not represent the center well at all. What about values like 81.5, 86, 89.6, 93, or 115? Which of those, if any, would be the best number to represent the center?

   Write down your choice for the number that best represents the center of all ten book prices: ______________

3. What does it mean to be the center of a set of data? We could develop lots of definitions (like I did when I calculated options d-h on the previous page), but how would we decide which definition is best?

We could come up with criteria to decide which number best represents the center. One criterion that students typically think of is this: *The best center is the number that is closest to all the data values.*

Another way of stating this criterion would be: *The best center is the number that minimizes the distances between itself and all the data values.*

We could quantify that criterion with: $\min\left\{\sum_{i=1}^{n}(x_i - c)\right\}$. The best center (c) minimizes the sum of the distances from itself to all the data values.

Let's try substituting a value for *c* into our criterion. For now, I'll arbitrarily choose c = 100 to represent my guess at the best center.

The table to the right shows how I calculated a value of -70 for this criterion (using c = 100). The best center, would minimize this criterion, so I could try another value (other than 100) and see what I get.

Data (x)	Center (c)	Criterion: (x – c)
3	100	-97
4	100	-96
10	100	-90
10	100	-90
11	100	-89
152	100	52
153	100	53
180	100	80
180	100	80
227	100	127
	Sum:	**-70**

… uh oh. I see a problem. What does it mean to minimize something that can become negative? If we used a crazy value for c like c = 1000000, we're guaranteed to get a criterion value less than -70.

Also, notice that far right column in the table. When our center is greater than a data value, we calculate a negative distance. When our center is to the right of a data value, we calculate a positive distance. All these negative and positive distances are going to cancel each other out (at least slightly). We need some way to make sure all our distances are positive.

Maybe we can modify our criterion to use the absolute value of the distance between the center and each data value. That would ensure we get positive distances. Let's try it.

Our new criterion, which we could call the *minimum absolute distance criterion* would be: $\min\left\{\sum_{i=1}^{n}|x_i - c|\right\}$

Let's see what we get if we try out c = 100 as our best guess for the center.

Data (x)	Center (c)	Criterion: |x – c|
3	100	97
4	100	96
10	100	90
10	100	90
11	100	89
152	100	52
153	100	53
180	100	80
180	100	80
227	100	127
	Sum:	**854**

The table on the left shows the sum of our absolute distances is 854 when we use 100 as our center.

If we try a different center, we'll get a different sum. Whichever center gives us the smallest sum possible is, by our definition, the best center.

On the top of the next page, I try out some other possible values for the center. You will be asked to identify which of those three values represents the best center:

Data (x)	Center (c)	|x – c|
3	8	5
4	8	4
10	8	2
10	8	2
11	8	3
152	8	144
153	8	145
180	8	172
180	8	172
227	8	219
	Sum:	**868**

Data (x)	Center (c)	|x – c|
3	14	11
4	14	10
10	14	4
10	14	4
11	14	3
152	14	138
153	14	139
180	14	166
180	14	166
227	14	213
	Sum:	**854**

Data (x)	Center (c)	|x – c|
3	80	77
4	80	76
10	80	70
10	80	70
11	80	69
152	80	72
153	80	73
180	80	100
180	80	100
227	80	147
	Sum:	**854**

Of the four values I arbitrarily chose for the "best" center, circle the one that was best: 8 14 80 100

4. It looks like we still have a problem. Using our criterion, there are multiple values that can all claim to be the "best" center of this data. Let's try to find the value of the center that will give us the *true minimum* of those absolute distances.

   For now, we'll find this minimum graphically. You could do this on a graphing calculator, but it might worthwhile to introduce you to a free online graphing calculator at desmos.com.

   a) Go to http://www.desmos.com and click **LAUNCH CALCULATOR**

   b) On the left, you can enter functions that will be graphed to the right. Go ahead and try typing **y = 2x + 4**

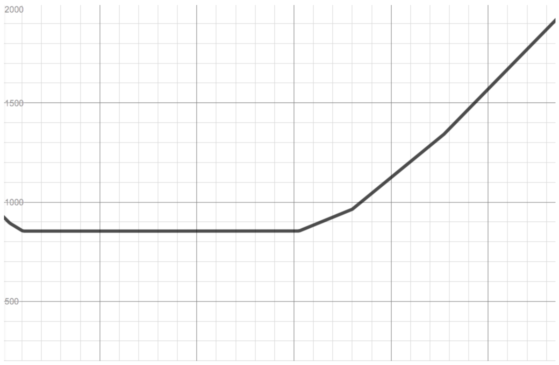
   c) We want to graph our criterion function. Remember, our criterion finds the sum of the absolute distances. With the data from this example, the function you need to type is:

   $$\mathbf{y = |3 - x| + |4 - x| + |10 - x| + |10 - x| + |11 - x| + |152 - x| + |153 - x| + |180 - x| + |180 - x| + |227 - x|}$$

   d) Once you type this in, you'll notice… nothing. We need to set the scales for the x- and y-axes to get the graph of our function to appear.

   To change the scales of the axes, click the 🔧 icon at the top-right of the screen. That will open a pop-up menu that will allow you to type in new limits for your x- and y-axes. We already know our data range from 3 to 227, so we can set the x-axis to go from 0 to 300. The y-axis is trickier to set. For now, go ahead and change it so that y ranges from 0 to 2000. You should get a graph like the one displayed below:

   e) Now, we simply have the find the value(s) of x that minimize this function. I've hidden the x-axis on the graph to the right, but you can see it on Desmos. For what values of X is our function minimized? To find more exact values, you can click on your line and Desmos will give you the coordinates.



   Answer: ______________________________

5. It's not very satisfying to have multiple values be the "best" center, is it? Ideally, we'd have one single value. Look at the end points from your previous answer (the lowest and highest possible values of the "best" center). Take those values and find their midpoint.

   That midpoint should match one of the possible centers I provided on the first page of this assignment.

   What value did you get for the midpoint? ______________ What do we call this midpoint? __________________

6. At this point, you might be wondering why the mean (average) isn't the best center. If it isn't the best, why were we all taught to calculate averages all our lives?

   Remember earlier (question #3) I wrote that we could choose multiple definitions and criteria to define the best center. The sum of the absolute distances gave us the value you just wrote for your answer to question #5. If we use a different criterion, we'll get a different value for the best center.

   Also remember why we used absolute values in our criterion: we needed to make sure all our distances were positive. We didn't have to use absolute values, but it seemed like the most straightforward way to ensure all the distances were positive.

   There is another simple way to make all numbers positive. Whenever we square a number, it's guaranteed to be positive. What would happen if we use squared distances instead of absolute distances in our criterion?

   Not only would we get positive distances, we'd get much larger distances when our center was far from a particular data value. So this criterion is sensitive to outliers. If we have values in our data that are extremely large or small, they will be far away from our center. By squaring that distance, we are making it even larger.

   Our criterion, which we could call the *minimum squared distance criterion* would then be: $\min\left\{\sum_{i=1}^{n}\left(x_i - c\right)^2\right\}$

   We could calculate this value for various values of c to see which center is better than the others:

Data (x)	Center (c)	(x-c)
3	14	121
4	14	100
10	14	16
10	14	16
11	14	9
152	14	19044
153	14	19321
180	14	27556
180	14	27556
227	14	45369
	Sum:	**139108**

Data (x)	Center (c)	(x-c)
3	80	5929
4	80	5776
10	80	4900
10	80	4900
11	80	4761
152	80	5184
153	80	5329
180	80	10000
180	80	10000
227	80	21609
	Sum:	**78388**

Data (x)	Center (c)	(x-c)
3	100	9409
4	100	9216
10	100	8100
10	100	8100
11	100	7921
152	100	2704
153	100	2809
180	100	6400
180	100	6400
227	100	16129
	Sum:	**77188**

   Of the three values I arbitrarily chose for the "best" center, circle the one that was best: 14 80 100

7. So we found which value for the center was better than the other two, but we still haven't found the "best" center with this new criterion.

To find the value of c that minimizes our criterion, we could turn again to Desmos.

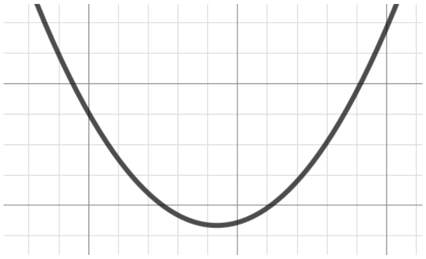
a) Go to http://www.desmos.com and click **LAUNCH CALCULATOR**

c) Type in our criterion function (now with squared distances). I won't type it all out here, but you will need to type the entire function into Desmos.

$y = (3 - x)^2 + (4 - x)^2 + (10 - x)^2 + \dots + (227 - x)^2$

d) Set the axes. The x-axis can, once again, range from 0 to 300. The y-axis will need to be much larger.

e) You'll get a graph similar to the one on the right. I've hidden the x- and y-axis labels, but you can see that it does appear as though this function has a single minimum. To find the coordinates of that minimum, click near the minimum and Desmos will help you out.



What value of c minimizes this function? ______________________.

Look again at the values I calculated on page one. What do we call this "best" center? ______________________.

8. So it looks like the mean and median can both stake claim as being the best center for a dataset. The median minimizes the sum of the absolute distances, while the mean minimizes the sum of the squared distances.

If I had to choose one, which would be the absolute best center for the data set in this assignment? Neither one.

Take another look at our data (the dotplot on page one). Now, look at where the median and mean would be on that dotplot. Neither the mean nor the median are very close to our data, so neither of them are very representative of our data. So while we can always calculate the mean and median of a set of data, those center values might not be very meaningful or useful.

Finally, let's take a look at a quick derivation of the mean. First, I write out a general form of our criterion (with a representing the best center). Then, by setting the derivative equal to zero, I can minimize the function to find that the best center is what we get when we add up all our data values and divide by the number of values we have.

$$\sum_{i=1}^{N}(x_i - a)^2 = (x_1 - a)^2 + (x_2 - a)^2 + \dots + (x_n - a)^2 = \left(x_1^2 - 2x_1a + a^2\right) + \left(x_2^2 - 2x_2a + a^2\right) + \dots\left(x_n^2 - 2x_na + a^2\right)$$

$$\frac{d}{dx} = (2x_1 - 2a) + (2x_2 - 2a) + \dots + (2x_n - 2a) = 2(x_1 + x_2 + \dots + x_n) - 2(a + a + \dots + a)$$

$$0 = 2(x_1 + x_2 + \dots + x_n) - 2(a + a + \dots + a) \Rightarrow (x_1 + x_2 + \dots + x_n) = (a + a + \dots + a) = na$$

$$\frac{(x_1 + x_2 + \dots + x_n)}{n} = a = \frac{1}{n}\sum x_i$$