Measures of center and spread

1 Measures of center and spread

Last time, we discussed the center and spread of a distribution through graphs. In this chapter, we'll see how to measure the center and spread with numbers.

2 Mean as a measure of center

The most commonly used measure of center is the arithmetic mean, or simply called mean. The mean is also known as the average of the set of data values. Mean is calculated as

$$\bar{x} = \frac{\Sigma x}{n}$$

where $\bar{x}$ is the sample mean, Σ (sigma) represents the sum of all the values in the dataset, and n is the total number of observations in your sample.

Example: Assume that these are the exam scores of 10 students in an introductory statistics course: 65,67,72,74,75,75,78,81,82,83. Then the mean is calculated as:

$$65 + 67 + 72 + 74 + 75 + 75 + 78 + 81 + 82 + 83 = \frac{752}{10} = 75.2 \implies \bar{x} = 75.2$$

So, the average score in an exam was 75.2. Sounds reasonable!

Now let's include outliers. For the same dataset as above, let's change the two lowest scores of 65 and 67 to a 5 and a 10 and calculate the average exam scores of 10 students.

$$\bar{x} = \frac{5 + 10 + 72 + 74 + 75 + 75 + 78 + 81 + 82 + 83}{10} = 56$$

This suggests the average class score was 56. It doesn't sound reasonable.
Reason - Outliers will skew the mean.

Now let's assume exam was a very difficult exam and these are the scores of 18 students in that exam: $10, 13, 18, 17, 22, 23, 25, 29, 30, 31, 32, 45, 65, 72, 78, 81, 89, 94$. Now let's calculate the average score:

$$\bar{x} = \frac{10 + 13 + 18 + 17 + 22 + 23 + 25 + 29 + 30 + 31 + 32 + 45 + 65 + 72 + 78 + 81 + 89 + 94}{18} = 43$$

This suggests the average class score was 43. It doesn't sound reasonable either.
Reason - The dataset was a right skewed dataset. Mean is not the good representation of center in a skewed dataset.

Question: But why did the mean work in the first example then? **Reason - Mean is a good representation of center only in a symmetric dataset.**

Now, we need to find an alternative metric to measure the center that is robust to skewed datasets and outliers.

3 Median as a measure of center

The median of a sample of data cuts a distribution down the middle, so about 50% of the observations are below it, and about 50% are above it when the dataset is sorted in an ascending order (lowest to highest). To calculate the value of the median, follow these steps:

- Sort the data from smallest to largest.
- If the set contains an odd number of observed values, the median is the middle observed value.
- If the set contains an even number of observed values, the median is the average of the two middle observed values.

Example: The median of $1, 2, 3, 4, 5$ is 3 as 3 is the middle number. 1 and 2 are to the left, and 4 and 5 are to the right.

To calculate the median of $1, 2, 3, 4, 5, 6$ we do not have a middle number. Both 3 and 4 are in the middle. Hence the median will be average of 3 and 4 which is 3.5.

Now let's revisit the previous examples we used on mean.

Example: Assume that these are the exam scores of 10 students in an introductory statistics course: 65,67,72,74,75,75,78,81,82,83. Then the median is 75.

Sounds reasonable, and similar result to mean which as 75.2.

Case 2: when we include outliers by changing the two lowest scores of 65 and 67 to a 5 and a 10. The dataset looked like 5,10,72,74,75,75,78,81,82,83. The median is still 75!.

This suggests the median class score was 75. It sounds reasonable.
Observation - Median is robust to outliers.

Case 3:The exams scores in a difficult test: $10, 13, 18, 17, 22, 23, 25, 29, 30, 31, 32, 45, 65, 72, 78, 81, 89, 94$. Now let's calculate the median score which is 30.5.

This suggests the median class score was 43. It sounds reasonable.
Observation - Median better represents the center on a skewed dataset.

Question - What about symmetric dataset where both mean and median were roughly the same? In a symmetric dataset, we use mean to represent the center as it is more precise than median. In a perfectly symmetrical data, mean and median are equal.

4 Percentiles and quartiles

"Percentile" is in everyday use and is a number where a certain percentage of scores fall below that number. For example, you might might have seen "Your SAT score is 82nd percentile" which means 82% of scores fall below your SAT score.

Quartiles are special cases of percentiles. We usually study 25th percentile, 50th percentile, and 75th percentile. 25th percentile helps us determine the 25% point in the dataset, and is also called 1st Quartile (Q1). 50th percentile helps us determine the midpoint of dataset which is the median. 75th percentile helps us determine the 75% point in the dataset, and is also called 3rd Quartile (Q3).

How do we calculate the first quartile (Q1) and third quartile (Q3) then?
Firstly, we start by sorting the datset in ascending order. Then we calculate the median which gives us the halfway point of the dataset. Then we calculate the median of the first half which gives us the 25% point in the dataset i.e. Q1. Median of the 2nd half gives us the 75% point in the dataset i.e. Q3.

Calculate Q1, median, and Q3 of the following dataset: 2, 10, 11, 12, 15, 16, 18, 22, 23, 42, 53.

Since there are 11 data points, median is the 6th item which is 16. Then, we just calculate median of the first half: 2, 10, 11, 12, 15 which is 11. Thus, Q1 = 11. Similarly, we just calculate median of the second half: 16, 18, 22, 23, 42, 53 which is 22. Thus, Q3 = 22.

Calculate Q1, median, and Q3 of the following dataset: 65,67,72,74,75,76,78,81,82,83.

Since there are 10 data points, median is the average of 5th and 6th item which is 75.5. Then, we just calculate median of the first half (whatever is to the left of 75.5): 65, 67, 72, 74, 75 which is 72. Thus, Q1 = 72. Similarly, we just calculate median of the second half (whatever is to the right of 75.5): 76, 78, 81, 82, 83 which is 81. Thus, Q3 = 81.

Note: This is just one way of calculating Q1 and Q3. Different softwares use different methods to calculate the, so the answers might differ.

5 Measures of spread/variability

The spread of a distribution describes how close the data values are to each other. When the spread of the distribution is combined with a measure of center, a good description of the data set is given.

As we established before that we should use mean to describe the center in a symmetric dataset and median in a skewed dataset, we have two different metrics to measure variability as well.

- Standard deviation

Standard deviation measures the variability around the mean and represents the average amount of variability in your dataset. **It tells you, on average, how far each value lies from the mean.** Small standard deviation indicates data are clustered tightly around the mean, and large standard deviation indicates data are more spread out.

How to calculate sample standard deviation?

Sample standard deviation (s) is calculated using the formula below where $\bar{x}$ represents the mean, and N is the number of observations.

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{N - 1}}$$

You can follow the following steps to calculate standard deviation.

1. Find the mean.
2. For each data point, find the square of its distance to the mean.
3. Sum the values from Step 2.
4. Divide by the number of data points.
5. Take the square root.

Example: Calculate the standard deviation of the following dataset: 1, 2, 3, 4, 5.

Firstly we calculate the mean of the dataset. $\bar{x} = \frac{1+2+3+4+5}{5} = 3$

x	$x - \bar{x}$	$(x - \bar{x})^2$
1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

$\sum(x - \bar{x})^2 = 10$. Therefore, $s = \sqrt{\frac{10}{5-1}} = \sqrt{\frac{10}{4}} = 1.58$

So this means on the dataset we had, on average, values is 1.58 away from mean.

A simple example to visualize: Calculate the standard deviation of the following dataset: 0, 0, 44, 88, 88

Firstly we calculate the mean of the dataset. $\bar{x} = \frac{0+0+44+88+88}{5} = 44$

x	$x - \bar{x}$	$(x - \bar{x})^2$
0	-44	1936
0	-44	1936
44	0	0
88	44	1936
88	44	1936

$\sum(x - \bar{x})^2 = 7744$. Therefore, $s = \sqrt{\frac{7744}{4}} = 44$

So this means on the dataset we had, on average the points are 44 away from mean. It totally makes sense!!

- Interquartile range (IQR)

Interquartile range mesaures the variability around the median and represents the average amount of variability in your dataset. **Interquartile range approximates the amount of spread in the middle half of the data.** Since median is robust to outliers, IQR is also roboust top outliers and is coupled with median to describe a dataset. IQR is calculated as

$$IQR = Q3 - Q1$$

Example: Calculate the IQR of the following dataset: 5, 17, 22, 24, 25, 36, 38, 41, 42, 53.

To find the IQR we firstly need to calculate Q1 and Q3. Since $Q1 = 22$ and $Q3 = 41$, $IQR = Q3 - Q1 = 41 - 22 = 19$.

Thus, we know that there is approximately a spread of 19 in the middle half of data.

Recap: We use mean and standard deviation to measure center and spread on a symmetric dataset, and we use median and IQR to measure center and spread on a skewed dataset.

6 Outliers

In previous chapter we described outliers as the extremely high or extremely low data values that do not follow the trend of the rest of the day. Now we know how to calculate the spread of a dataset, we can formally define a threshold for outliers.

- Using Standard deviation

It is true that we use standard deviation to calculate spread on a symmetric dataset, and it is also a natural observation is standard deviation itself is affected by outliers. Hence, there is not a concrete rule to detect outliers using standard deviation. On a perfectly symmetrical dataset, it is said that datapoints above or below 3 standard deviation of mean are outliers. However, on a roughly symmetrical dataset, data-points above or below 2 standard deviation of mean can be considered as outliers.

- Using IQR

On a skewed dataset where we use IQR to define a spread, we have a concrete rule to detect outliers. Any observations that are more than 1.5 interquartile ranges below the first quartile or above the third quartile are outliers i.e. any observations below $Q1 - 1.5 \times IQR$ or any observations above $Q3 + 1.5 \times IQR$ are outliers.

Example: For the data set: 45 , 58, 59, 62, 64, 66, 75 decide if the values 45 and 75 are considered outliers or not.

First step to detect outliers is to find IQR. To find IQR we need to find Q1 and Q3. To do so, we need to find median. Since there are 7 observations, median is the 4th observation which is 62. Median of the first half: 45 , 58, 59 is Q1 which is 58. Median of the second half: 64, 66, 75 is Q3 which is 66.

Now, $IQR = Q3 - Q1 = 66 - 58 = 8$.

Lower fence $= Q1 - 1.5 \times IQR = 58 - 1.5 \times 8 = 46$.

Upper fence $= Q3 + 1.5 \times IQR = 66 + 1.5 \times 8 = 78$.

Hence, 45 is an outlier, and 75 is not an outlier.

7 What's greater, mean or median?

We established above that mean and median are roughly the same for a symmetric dataset. What about skewed dataset? On a skewed dataset because of outliers, we saw that mean will be towards the center of the dataset.

On a right skewed dataset we have lots of observations to the left of the graph hence median will be towards the left of the graph. Meanwhile mean is pulled towards the center by outliers and skeweness. Hence on a right skewed dataset, Mean > Median.

On a left skewed dataset we have lots of observations to the right of the graph hence median will be towards the right of the graph. Meanwhile mean is pushed towards the center by outliers and skeweness. Hence on a left skewed dataset, Median > Mean.