Measures of Variability

EDP 613

Week 4

Before we Begin

Remember that a statistic is **resistant** if its value is not affected by extreme values (large or small) in the data set. So

Q: Which of the measures of central tendency are resistant?

A: Since the *median* is simply the middle value, it is not affected by outliers and is therefore resistant.

Basic Idea

Variability basically tells us how far apart data points lie from each other and from the center of a distribution

Why?

Generally

The ***central tendency*** tells us where most of our points lie

The ***variability*** summarizes how far apart the points are

What Does it Tell Us?

Measures of Variability

Range

Interquartile range

Standard deviation

Variance

The Range

The ***range*** of a data set is the difference between the largest value (Max) and the smallest value (Min)

$$\text{range} = \text{Max} - \text{Min}$$

Example

Compute the **range** for the **sample** of people

4	1	1	3	4	7

While not necessary, putting the data set in numerical order reduces the likelihood of making a silly mistake

1	1	3	4	4	7

Steps

We have $\text{Max} = 7$ and $\text{Min} = 1$ so

$$7 - 1 = 6$$

or in context **6 people**

Example

Compute the **range** for the **sample** \$3.61, \$3.84, \$3.79, \$3.61, \$4.09, and \$3.96.

First for simplicity, we arrange the data set in numerical order

3.61	3.61	3.79	3.84	3.96	4.09

Steps

$$4.09 - 3.61 = 0.48$$

or in context **$0.48**

The interquartile range

Every data set has three quartiles

> - Q_1
>
>   > - first quartile
>   > - 25th percentile
>   > - separates the lower 25% of the data from the higher 75%
>
> - Q_2
>
>   > - second quartile
>   > - 50th percentile
>   > - separates the lower 50% of the data from the higher 50%%
>   > - aka the *median*
>
> - Q_3
>
>   > - third quartile
>   > - 75th percentile
>   > - separates the lower 75% of the data from the higher 25%

The *interquartile range* (IQR) is found by subtracting the first quartile from the third quartile

$$\text{IQR} = Q_3 - Q_1$$

Outliers

An *outlier* is a value that is considerably larger or smaller than most of the values in a data set

Finding Outliers: IRQ Method

1. Find the Min and Max
2. Find Q_1, Q_2, and Q_3
3. Compute the IQR
4. Compute the cutoff points for determining outliers - aka *outlier boundaries*

> Lower Outlier Boundary (LOB)
>
> $Q_1 - 1.5 \cdot \mathrm{IQR}$

> Upper Outlier Boundary (UOB)
>
> $Q_3 + 1.5 \cdot \mathrm{IQR}$

5. Any data point

> Less than the LOB
> is an outlier

> Greater than the UOB
> is an outlier

Example

Over the span of 35 days, Jamie drives to work every weekday morning and keeps track of her time (in minutes) for some reason

15	17	17	17	17	18	19
19	19	19	19	19	20	20
20	20	20	21	21	21	21
21	21	21	22	22	22	23
23	24	26	31	36	38	39

Construct a boxplot

Steps

1. We have
    - Max: **15 minutes**
    - Min: **39 minutes**

2. To find the position of Q_1, we have

$$\frac{25}{100} \cdot 35 = 0.25 \cdot 35$$

$$= 8.57$$

$$\approx 9$$

which tells to look for the data point in the 9th position

15	17	17	17	17	18	19
19	19	19	19	19	20	20
20	20	20	21	21	21	21
21	21	21	22	22	22	23
23	24	26	31	36	38	39

or in context **19 minutes**

Statistical Methods I

To find the position of Q_2, we have

$$\frac{50}{100} \cdot 35 = 0.50 \cdot 35$$

$$= 17.50$$

$$\approx 18$$

which tells to look for the data point in the 18th position

15	17	17	17	17	18	19
19	19	19	19	19	20	20
20	20	20	21	21	21	21
21	21	21	22	22	22	23
23	24	26	31	36	38	39

or in context the *median* is **21 minutes**

To find the position of Q_3, we have

$$
\begin{aligned}
\frac{75}{100} \cdot 35 &= 0.75 \cdot 35 \\
&= 26.25 \\
&\approx 26
\end{aligned}
$$

which tells to look for the data point in the 26th position

15	17	17	17	17	18	19
19	19	19	19	19	20	20
20	20	20	21	21	21	21
21	21	21	22	22	22	23
23	24	26	31	36	38	39

or in context **22 minutes**

3. To find the range between quartiles, we have

$$\begin{aligned}\text{IQR} &= 22 - 19 \\ &= 3\end{aligned}$$

or in context **3 minutes**

4. To find the boundaries, we have

$$\begin{aligned} \text{LOB} &= 19 - 1.5 \cdot 3 \\ &= 19 - 4.5 \\ &= 14.5 \end{aligned}$$

$$\begin{aligned} \text{UOB} &= 22 + 1.5 \cdot 3 \\ &= 22 + 3 \\ &= 26.5 \end{aligned}$$

giving us **14.5** and **26.5 minutes**, respectively

Statistical Methods I

Five-number summary

Report on

Min

Q_1

Q_2

Q_3

Max

Example

Following are the number of grams of carbohydrates in 12-ounce espresso beverages offered at Starbucks

14	43	38	44	31	27	39	59	9	10	54
14	25	26	9	46	30	24	41	26	27	14

First we will benefit from reordering the data set

9	9	10	14	14	14	24	25	26	26	27	27	30	31	38	39	41	43	44	46	54	59

Steps

1. We have
   - Min: **9 grams**
   - Max: **59 grams**

2. To find the position of Q_1, we have

$$\frac{25}{100} \cdot 22 = 0.25 \cdot 22$$
$$= 5.50$$
$$\approx 6$$

which tells to look for the data point in the 6th position

9	9	10	14	14	14	24	25	26	26	27	27	30	31	38	39	41	43	44	46	54	59

or in context **14 grams**

To find the position of Q_2, we have

$$\frac{50}{100} \cdot 22 = 0.50 \cdot 22$$
$$= 11$$

which tells to look for the data point in the 11th position

9	9	10	14	14	14	24	25	26	26	27	27	30	31	38	39	41	43	44	46	54	59

or in context the *median* is **27 grams**

To find the position of Q_3, we have

$$\frac{75}{100} \cdot 22 = 0.75 \cdot 22$$
$$= 16.50$$
$$\approx 17$$

which tells to look for the data point in the 17th position

9	9	10	14	14	14	24	25	26	26	27	27	30	31	38	39	41	43	44	46	54	59

or in context **41 grams**

3. To find the range between quartiles, we have

$$\begin{aligned}\text{IQR} &= 41 - 14 \\ &= 27\end{aligned}$$

or in context **27 grams**

4. To find the boundaries, we have

$$
\begin{aligned}
\text{LOB} &= 14 - 1.5 \cdot 27 \\
&= 14 - 40.5 \\
&= -26.5
\end{aligned}
$$

$$
\begin{aligned}
\text{UOB} &= 41 + 1.5 \cdot 27 \\
&= 41 + 40.5 \\
&= 81.5
\end{aligned}
$$

giving us **-26.5** and **81.5 grams**, respectively

Realistically this is between 0 and 81.5 grams unless you can make a good argument that coffee can have negative grams of carbohydrates

The standard deviation

In a nutshell, a *standard deviation* is just a number we use to tell how measurements for a group of things are spread out from the average which in our case is the mean

Population

$$\sigma = \sqrt{\frac{\sum \left(Y - \bar{Y}\right)^2}{N}}$$

Sample

$$s = \sqrt{\frac{\sum \left(Y - \bar{Y}\right)^2}{n - 1}}$$

Y is a data point

$\bar{Y}$ is the mean

N is the **population** size

σ is the **population** **standard deviation**

n is the **sample** size

s is the **sample** **standard deviation**

If you want to know why we divide by `n-1` in a sample standard deviation, that is a pretty interesting topic and you can explore more about that over at Khan Academy

What Do These Look Like?

Example

Calculate the **sample** **standard deviation** of the following set of data points by hand

46	69	32	60	52	41

Again, putting the data set in numerical order can make it easier to track

32	41	46	52	60	69

Steps

1. Compute the mean

$$\bar{Y} = \frac{32 + 41 + 46 + 52 + 60 + 69}{6}$$

$$= \frac{300}{6}$$

$$= 50$$

2. Compute the deviations and square them

Y	$Y - \overline{Y}$	$\left(Y - \overline{Y}\right)^2$
32	-18	324
41	-9	81
46	-4	16
52	2	4
60	10	100
69	19	361

3. Calculate the sum of (the) squares

$$\left(Y - \overline{Y}\right)^2 = 324 + 81 + 16 + 4 + 100 + 361$$

$$= 886$$

4. Divide by size

$$\frac{886}{6-1} = \frac{886}{5}$$

$$= 177.2$$

5. Take the square root

$$\sqrt{177.2} \approx 13.31$$

implying that ***each data point deviates from the mean by 13.31 points on average***

The Variance

In a nutshell, a *variance* is just a number we use to tell how measurements for a group of things are spread out from the average which in our case is the mean and the measure is always positive

Population

$$\sigma^2 = \frac{\sum\left(Y - \overline{Y}\right)^2}{N}$$

Sample

$$s^2 = \frac{\sum\left(Y - \overline{Y}\right)^2}{n-1}$$

Y is a data point

$\overline{Y}$ is the mean

N is the **population** size

σ is the **population variance**

n is the **sample** size

s is the **sample variance**

Example

Calculate the **variance** of the following set of data points by hand

46	69	32	60	52	41

We actually already calculated this! Let's go back to step 4

4. Divide by size

$$\frac{886}{6-1} = \frac{886}{5}$$

$$= 177.2$$

This is actually the **sample variance**

Joined at the Hip

The **standard deviation** is just the square root of the **variance**

or equivalently

the **variance** is just the square of the **standard deviation**

so

you can't have one without the other

That's it. Let's take a break before working in R.