Summarizing the Center and Spread of Quantitative Variables

September 17, 2018

Summaries of Center (what is a "typical" value?)

Reminder of definitions from your reading:

Suppose we observe n numbers, $x_1, \dots, x_n$.

There are two commonly used statistics used to summarize the **center** of the distribution of these values:

- The **mean** is the average of these values (add them up and divide by n). We use $\bar{x}$ to denote the mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \cdots + x_n}{n}$$

- The **median** is the middle value when you arrange them in order. (If the sample size n is even, you take the average of the middle two values)

The situation:

It's 2013, and 6 friends are hanging out at their local bar. Their incomes are \$30,000, \$32,000, \$34,000, \$36,000, \$38,000, and \$40,000.

What is their mean income?

What is their median income?

In walks BILL GATES!

According to the internet, in 2013 Bill Gates had an income of $11.5 billion (in case you're curious, that works out to $23,148 per minute).

What is the mean income of the 6 friends and Bill Gates? (Note that if you write it out with all the zeros, 11.5 billion is 11500000000; there are 8 zeros)

What is the median income of the 6 friends and Bill Gates?

Summaries of Spread (how spread out are the values?)

There are three common measures of the **spread** of a distribution (how “wide” is it?):

1. We saw the **inter-quartile range** (IQR) last class:

IQR = Q3 - Q1 = 75th percentile - 25th percentile

The IQR is the width of an interval covering the middle half of the data.

2. The **variance** is (almost) the average squared difference of each observation from the mean.

$$\text{Variance} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

3. The **standard deviation** is the square root of the variance. Intuitively, you can think of it as the average distance of the data points from the mean (although technically, that’s not exactly right).

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Let’s use R to calculate these, rather than doing it by hand.

I have set up two different data frames - one with the numbers for just the friends, and a second including both the friends and Bill Gates.

Here are summaries of center and spread, including just the friends:

```
friends
```

```
##   person_number salary
## 1             1  30000
## 2             2  32000
## 3             3  34000
## 4             4  36000
## 5             5  38000
## 6             6  40000
```

```
friends %>%
  summarize(
    mean_salary = mean(salary),
    median_salary = median(salary),
    iqr_salary = IQR(salary),
    var_salary = var(salary),
    sd_salary = sd(salary)
  )
```

```
##   mean_salary median_salary iqr_salary var_salary sd_salary
## 1       35000         35000       5000    1.4e+07  3741.657
```

Here are summaries of center and spread, including Bill too:

```
friends_and_bill
```

```
##   person_number   salary
## 1             1 3.00e+04
## 2             2 3.20e+04
## 3             3 3.40e+04
## 4             4 3.60e+04
## 5             5 3.80e+04
## 6             6 4.00e+04
## 7             7 1.15e+10
```

```
friends_and_bill %>%
  summarize(
    mean_salary = mean(salary),
    median_salary = median(salary),
    iqr_salary = IQR(salary),
    var_salary = var(salary),
    sd_salary = sd(salary)
  )
```

```
##   mean_salary median_salary iqr_salary   var_salary  sd_salary
## 1  1642887143         36000       6000 1.889274e+19 4346578211
```

What's the point?

Mean, Variance, and Standard deviation are sensitive to outliers and skewness. They should only be used when a distribution looks "nice" (unimodal, symmetric, no outliers). Otherwise, use median and IQR to summarize center and spread.

If the Distribution is...	Summarize Center with...	Summarize Spread with...
Unimodal, Symmetric, and no Outliers	mean most common; median also OK	standard deviation most common; variance or IQR also OK
Multimodal or Skewed or has Outliers	median	IQR