Summarizing Numeric Variables

February 5, 2018

Warm Up

Reminder of definitions from your reading:

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

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

The mean and median tell you about the **center** of a distribution.

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 the (almost) 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).

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

What is the IQR of the friends' incomes, not including Bill Gates?

You can find Q1 as the median of the values in the data set that are less than the median of the whole data set.

You can find Q3 as the median of the values in the data set that are greater than the median of the whole data set.

What is the variance of the friends' incomes, not including Bill Gates?

What is the standard deviation of the friends' incomes, not including Bill Gates?

I calculated these values for the friends' incomes including Bill Gates, and I got:

IQR = 6,000

Variance = 1.89e+19 = 1.89×10^{19} = 18900000000000000000

Standard Deviation = 4.35e+09 = 4.35×10^{9} = 4350000000

Do babies born to mothers who smoked during pregnancy weigh less on average than babies born to mothers who did not smoke?

Low birth weights are a risk factor for health problems later in life, so this is important!

Here is a data set with data on a sample of randomly selected babies who were born in North Carolina in 2004, with some information about the mother and the babies' weights in grams:

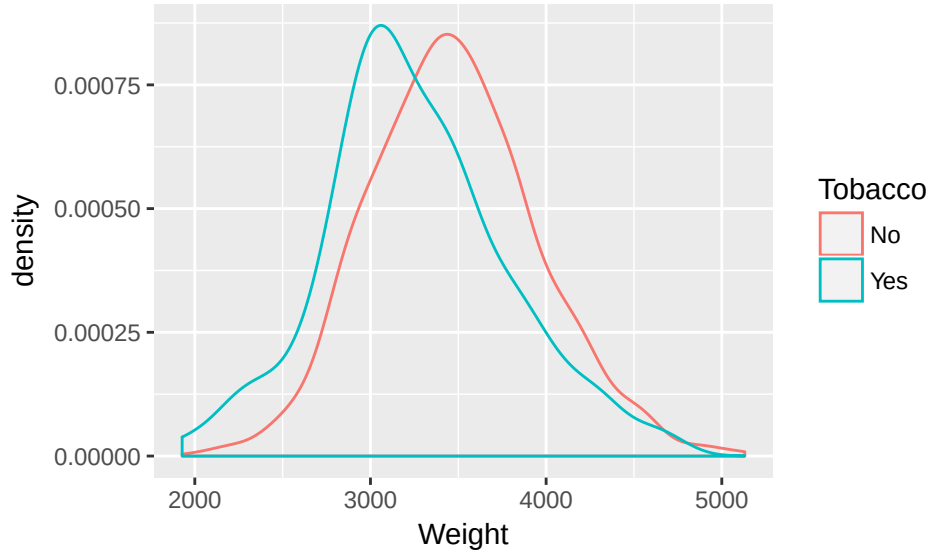
```
dim(NCBirths2004)
```

```
## [1] 1009    7
```

```
head(NCBirths2004)
```

```
##   ID MothersAge Tobacco Alcohol Gender Weight Gestation
## 1  1      30-34      No      No   Male   3827        40
## 2  2      30-34      No      No   Male   3629        38
## 3  3      35-39      No      No Female   3062        37
## 4  4      20-24      No      No Female   3430        39
## 5  5      25-29      No      No   Male   3827        38
## 6  6      35-39      No      No Female   3119        39
```

```
ggplot(data = NCBirths2004, mapping = aes(x = Weight, color = Tobacco)) + geom_density()
```



```r
NCBirths2004 %>%
  group_by(Tobacco) %>%
  summarize(
    mean_wt = mean(Weight),
    median_wt = median(Weight),
    q1_wt = quantile(Weight, probs = 0.25),
    q3_wt = quantile(Weight, probs = 0.75),
    iqr_wt = IQR(Weight),
    var_wt = var(Weight),
    sd_wt = sd(Weight)
  )
```

```
## # A tibble: 2 x 8
##   Tobacco  mean_wt median_wt q1_wt  q3_wt iqr_wt   var_wt    sd_wt
##    <fctr>    <dbl>     <dbl> <dbl>  <dbl>  <dbl>    <dbl>    <dbl>
## 1      No 3471.912      3459  3147 3771.0  624.0 229012.4 478.5524
## 2     Yes 3256.910      3204  2948 3529.5  581.5 270898.2 520.4788
```

Which of these statistics can we use to summarize the distributions of baby weights in each group?

Is there evidence that a mother's tobacco usage affects the baby's weight?