Unit2

Goals

- Summarize single variable real-valued data
- First data set
- R and Google Spreadsheet to compute mean

Datasets

Income data from the census http://www.census.gov/census2000/PUMS.html

Artificial height data

Summary statistics of a single variable

In this basic scenario, we are given a bunch of numbers (a bunch of observations), and they look like a bunch of numbers. The most simple way to make sense of it is to compute summary statistics.

A summary statistic is a number that summarize the data as simply as possible. We will learn the four main summary statistics. You might have learned these in high school, but I'm confident that there are subtleties about them that you might not have heard about.

Mean (Arithmetic mean)

Mean is the sum of all observations divided by the number of observations. It is supposed to tell the center or the middle of the data (formally "central tendency"). We also call this arithmetic mean because there are other kinds of means, but they are not as widely used as arithmetic mean.

Suppose we have collected the height data from some people in the room. And the data look like this (in centimeters)

$$175, 170, 180, 185, 165$$

Then, the arithmetic mean is $\frac{175+170+180+185+165}{5} = 175$ centimeters.

R Exercise

Simple enough right? Let's use R to compute the mean.

```
(175+170+180+185+165) / 5
```

```
## [1] 175
```

Alternatively, we can use `mean` function on a vector.

First, we need to enter our data into a vector.

```
heights = c(175, 170, 180, 185, 165)
heights
```

```
## [1] 175 170 180 185 165
```

Then, we compute the mean from the vector.

```
mean(heights)
```

```
## [1] 175
```

Formal definition

In general, the arithmetic mean is defined as follows:

Suppose we have data points $x_1, x_2, x_3, \ldots, x_n$. Then the arithmetic mean $\bar{x}$ of these data points are

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

The ability to understand mathematical notation can come in handy. You should note what the big sigma sign $\sum$, the $i = 1$, and the n mean. You will see the big sigma sign over and over.

Mean maximizes happiness.

This can be proven mathematically, but we will work through an example to demonstrate this point. Let's use the same dataset as before.

```
heights = c(175, 170, 180, 185, 165)
```

Suppose the torso length is 30% of the height. What should be the most appropriate torso length for each person?

```
torso_lengths = heights * 30 / 100
torso_lengths
```

```
## [1] 52.5 51.0 54.0 55.5 49.5
```

But we can only afford to have one size to fit all five people here. Let's try 51 inches and see how well they fit people.

```
error = torso_lengths - 51
error
```

```
## [1]  1.5  0.0  3.0  4.5 -1.5
```

The error represents how much off the shirt is. Some textbooks call this 'deviation', 'error', or 'residual'. But we will call it error. We cannot just sum it up yet. Why?

```
sum(error)
```

```
## [1] 7.5
```

Some of the errors will cancel out. It's like the errors from the first and the last persons don't count at all. We have to work around it. Squaring it seems like a fair choice.

```
error^2
```

```
## [1]  2.25  0.00  9.00 20.25  2.25
```

Now the first person and the last person contribute equally when we sum up all of these squared errors. Let's roll with this.

```
sum(error^2)
```

```
## [1] 33.75
```

Pretty awful, right? It's way too big for some people and way too small for some people. In fact, the mean is the best size i.e. it pleases most of the people overall.

```r
mean(torso_lengths)
```

```
## [1] 52.5
```

```r
error = torso_lengths - mean(torso_lengths)
error
```

```
## [1]  0.0 -1.5  1.5  3.0 -3.0
```

```r
sum(error ^ 2)
```

```
## [1] 22.5
```

The squared error is much lower than we fix the torso length of the shirt to 51 inches. You cannot do better than this because mean minimizes the sum of squared error. This is also why mass produced stuff is always meh. The factory always goes for the mean to please the most people.

To repeat, **arithmetic mean pleases the most people**.

Other examples :

- How long should the hand dryer go?
- How long should the automatic water faucet go?
- How sweet should the mass produced ice-cream be?

Exercise

Repeat the analysis on Google Spreadsheet

Work with real data

As we said before, a data table comes in a machine-readable file. More specifically, they might come in one of these formats.

1. Comma-separated file (.csv)

   ```
   height.foot, height.inch, weight.kg
   5, 9, 76
   6, 0, 88
   ```

2. Tab-separated file (.tsv)

   ```
   height.foot  height.inch    weight.kg
   5   9   76
   6   0   88
   ```

3. Excel file (.xls or .xlsx)

We will only use csv files in this class because it can be read by both R and Spreadsheet. Spreadsheet can read most formats. But csv is arguably the most convenient (and possibly standard)

Data frames and data import in R

First, download the dataset from here.

```
data = read.csv('datasets/heights.csv')
```

Explore the dataset

```
head(data)
```

```
##   weight height.foot height.inch
## 1     78           5          11
## 2     83           6           5
## 3     80           5          11
## 4     72           5           8
## 5     75           5           8
## 6     74           5          11
```

```
head(data$height.foot)
```

```
## [1] 5 6 5 5 5 5
```

```
head(data$weight)
```

```
## [1] 78 83 80 72 75 74
```

Transform and convert the data

```
data$height.cm = (data$height.foot  + data$height.inch / 12) * 30
```

Compute the mean

```
mean(data$height.cm)
```

```
## [1] 169.8
```

Bonus: how many people are at least 6 foot tall?

```
sum(data$height.foot >= 6)
```

```
## [1] 95
```

Bonus: what's the proportion of people who are at least 6 foot tall? There are two ways to compute this. Both are easy.

```
sum(data$height.foot >= 6) / length(data$height.foot)
```

```
## [1] 0.095
```

```
mean(data$height.foot >= 6)
```

```
## [1] 0.095
```

Exercise:

- Compute the mean weight in pounds.
- Compute the number of people who are shorter than 150 centimeters.
- Repeat the analysis on a Spreadsheet