Introduction to Statistics, Examples

CSU, Chico Math 314

2018-08-27

outline

Population

To which population does the following refer?

- Smoking causes lung cancer?
- Do Gmail users receive (to their inbox) less spam than Outlook users?
- Does LibreOffice crash less often than Microsoft Word?

Sample

What would a random sample look like from the following?

- Smoking causes lung cancer?
- Do Gmail users receive (to their inbox) less spam than Outlook users?
- Does LibreOffice crash less often than Microsoft Word?

Experiments or Observational Study

What would it take to make the following experiment?

- Smoking causes lung cancer?
- Do Gmail users receive (to their inbox) less spam than Outlook users?
- Does LibreOffice crash less often than Microsoft Word?

outline

Lil' Analysis

Suppose we're attempting to measure computer performance. We'll use the unit millions of instructions per second (mips). We'll consider the data set named speed, which contains information on various systems mips.

```
speed <- read.csv("https://roualdes.us/data/speed.csv")
tail(speed)
```

```
##                    system   mips year
## 131           AMD FX-8350  97125 2012
## 132  Intel Core i7 3770K 106924 2012
## 133 Intel Core i7 3630QM 113093 2012
## 134  Intel Core i7 4770K 133740 2013
## 135  Intel Core i7 5960X 238310 2014
## 136       Raspberry Pi 2   4744 2014
##     cores
## 131     1
## 132     1
## 133     1
```

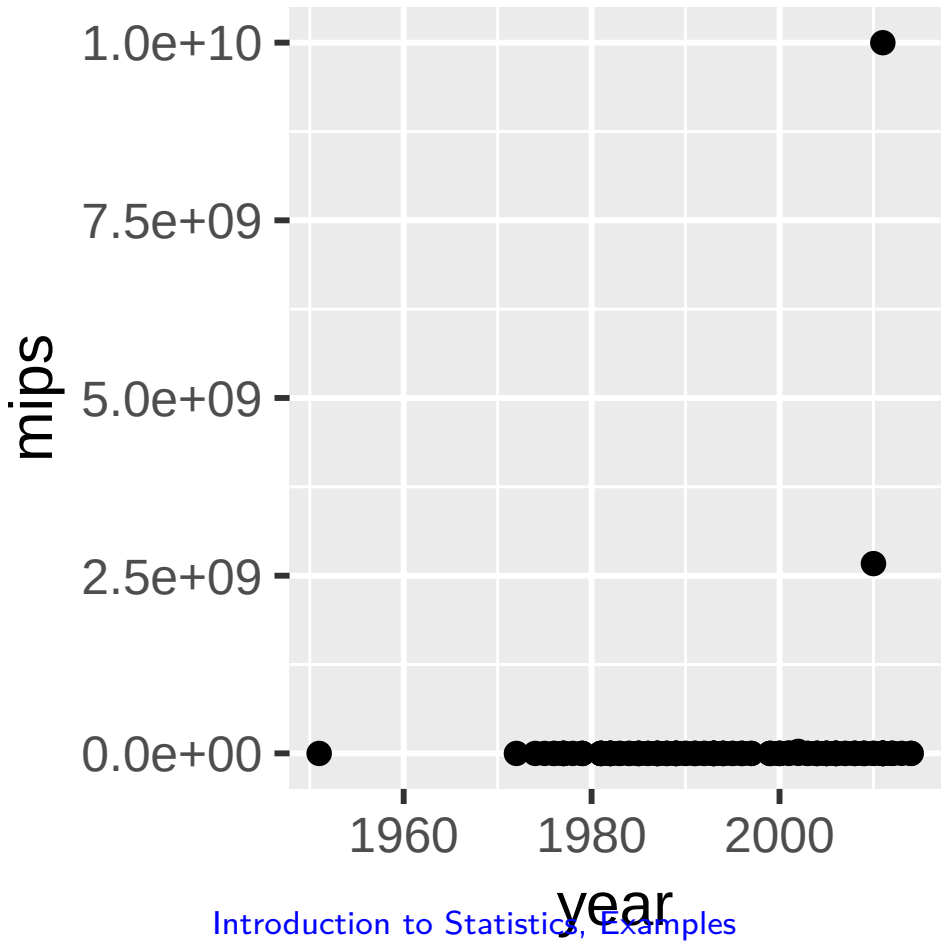
Lil' Analysis

Plot the data! What kind of data do we have? What kind of plot does that warrant?

Lil' Analysis

Make a scatter plot.

```
suppressMessages(library(ggplot2))
ggplot(data=speed, aes(year, mips)) +
    geom_point()
```

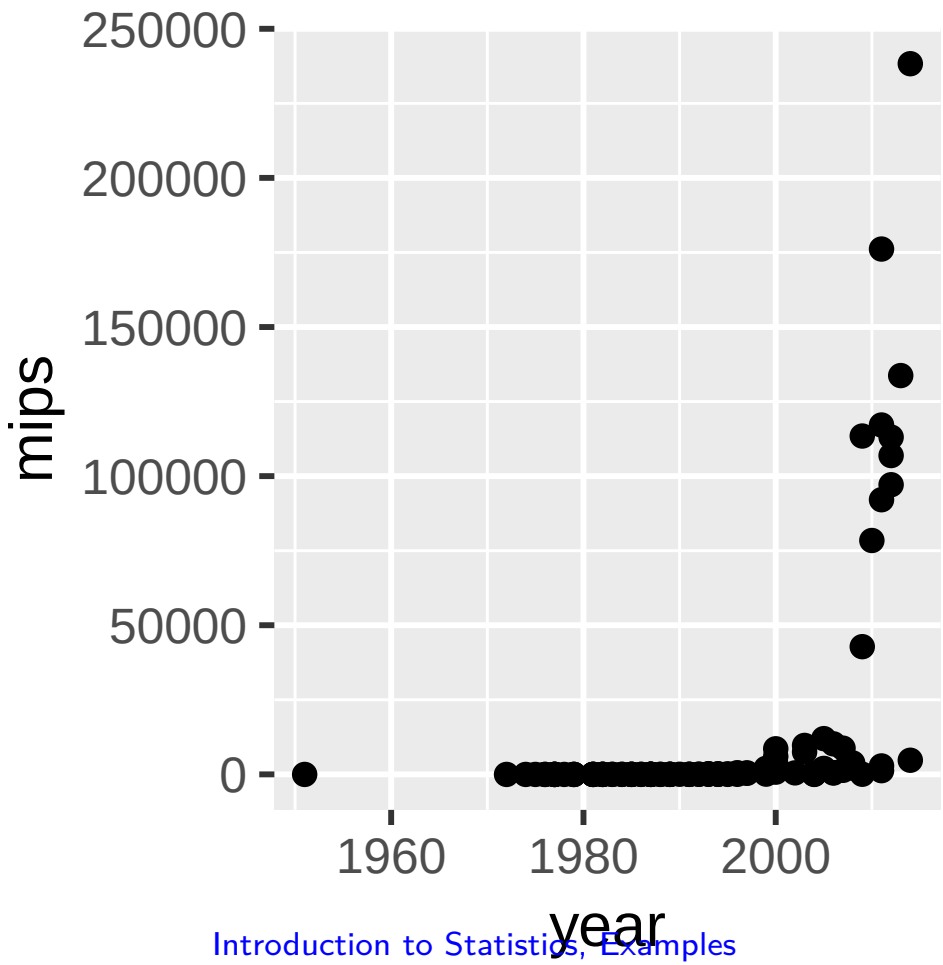


Lil' Analysis

What's wrong with this plot? Is there a lurking variable we are missing?

Lil' Analysis

```
suppressMessages(library(dplyr))
ggplot(data=dplyr::filter(speed, cores == 1),
       aes(year, mips)) +
    geom_point()
```



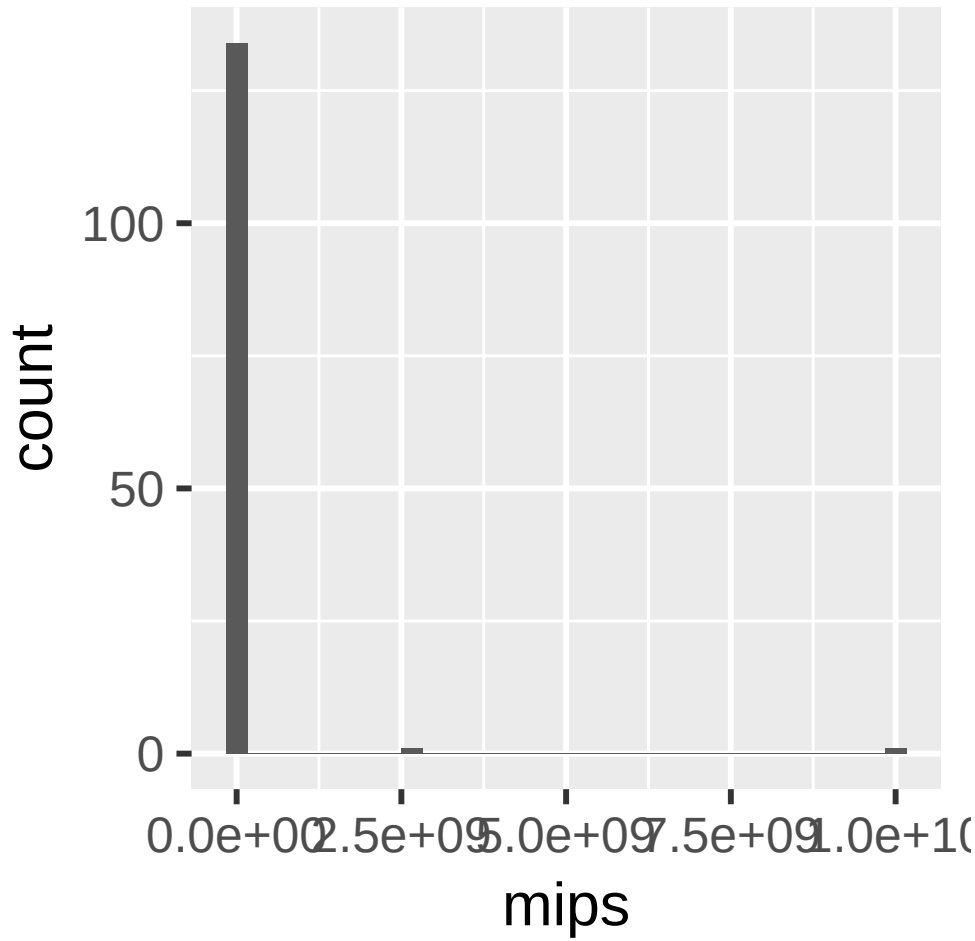
Lil' Analysis

From the above scatter plot, what kind of skew do we expect `mips` to have? What is the univariate plot that we should use to look at `mips`?

Lil' Analysis

Histogram of `mips`.

```
ggplot(aes(mips), data=speed) +
    geom_histogram(bins=31)
```



Lil' Analysis

Ignoring the variable `cores`, how can we find the mean and median?

Lil' Analysis

The functions `mean` and `median`.

```
mean(speed$mips)

## [1] 93427146

median(speed$mips)

## [1] 146.35
```

Lil' Analysis

Let's also calculate the standard deviation of `mips`.

```
sd(speed$mips)

## [1] 885866970

sd(filter(speed, cores == 1)$mips)

## [1] 39920.26
```

outline

Take Away

From a little analysis comes little conclusions.

- What can we conclude with respect to
  - sample and population?
  - plots?
  - mean, median, standard deviation?
  - What do we think of our simple models?
- What better questions might we be interested in?

outline

references I

Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. " O'Reilly Media, Inc.", 2016.