STAT 1291: Data Science

Lecture 3 - Data Visualization: What is a good graphic?

Sungkyu Jung

Where are we?

- *What is Data Science?*
- *How do we learn Data Science?*
  - We will use R and R Studio
  - This Friday's class will be at CL 244B. Bring your laptop with R installed.
- Data visualization

Understanding Univariate Data

Look at the **distribution** of univariate data

$$F(a) = \text{Prob}(\text{Height} \leq a)$$

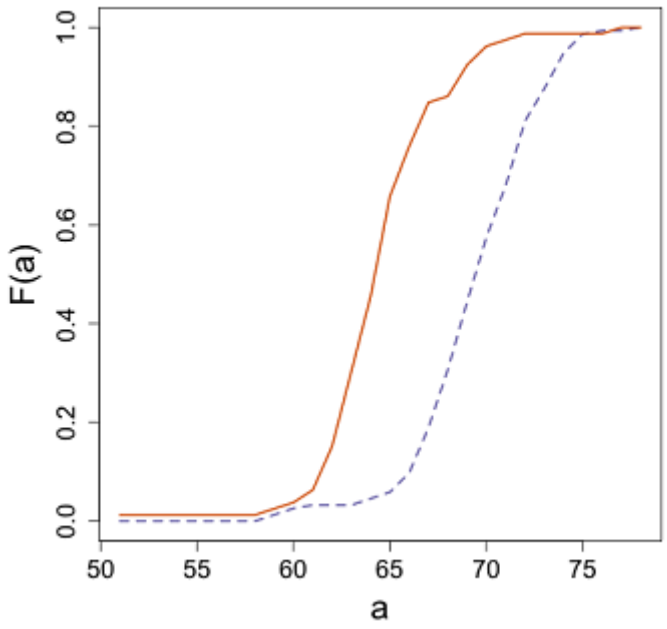

Figure 1:

Distributions

Histograms show: $F(b) - F(a)$ for several intervals $(a, b]$

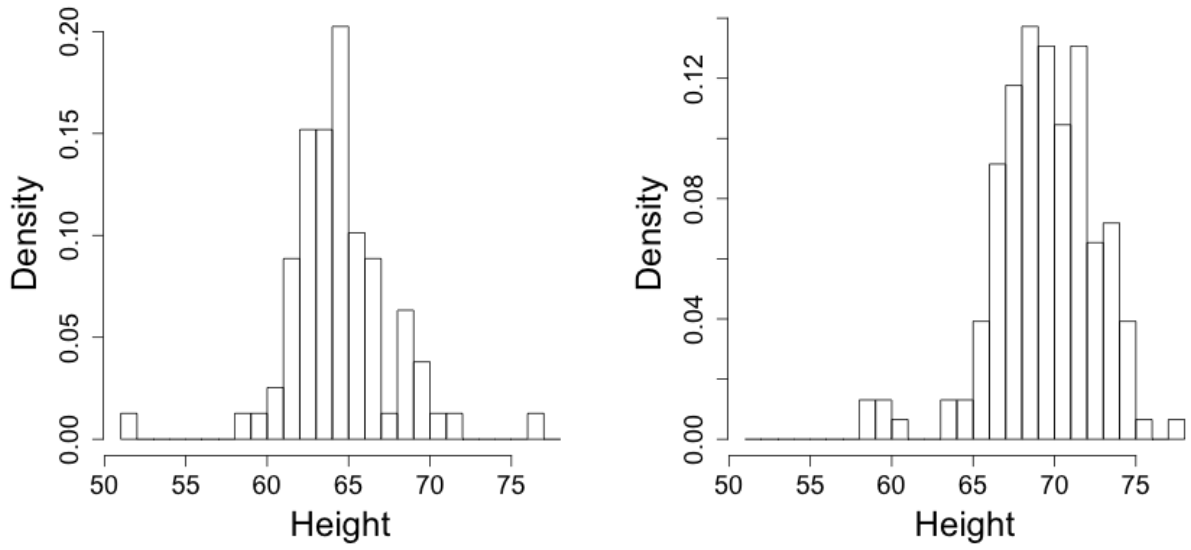

Figure 2:

Easier to interpret than cumulative distribution functions

Normal Approximation

The distribution of many outcomes in nature are approximated by the normal distribution:

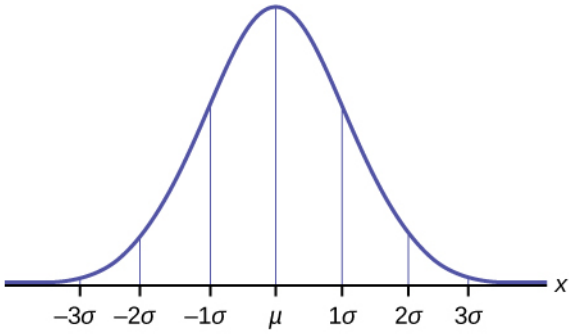

Figure 3:

- μ is the average (also called the mean)
- σ is the standard deviation

Normal Approximation

If our data follows the normal distribution then μ and σ are a sufficient summary: they tell us everything!

All we need to know is μ and σ

```
       Average SD
Male        70  3
Female      65  3
```

How good is the normal approximation?

Here are the approximations for males

```
  Height Real Approx
1     63 0.02   0.03
2     65 0.07   0.06
3     67 0.16   0.10
4     68 0.31   0.31
5     70 0.50   0.44
6     71 0.69   0.68
7     73 0.84   0.88
8     75 0.93   0.95
9     76 0.98   0.99
```

QQ-plots

Observed versus normal approximation quantiles

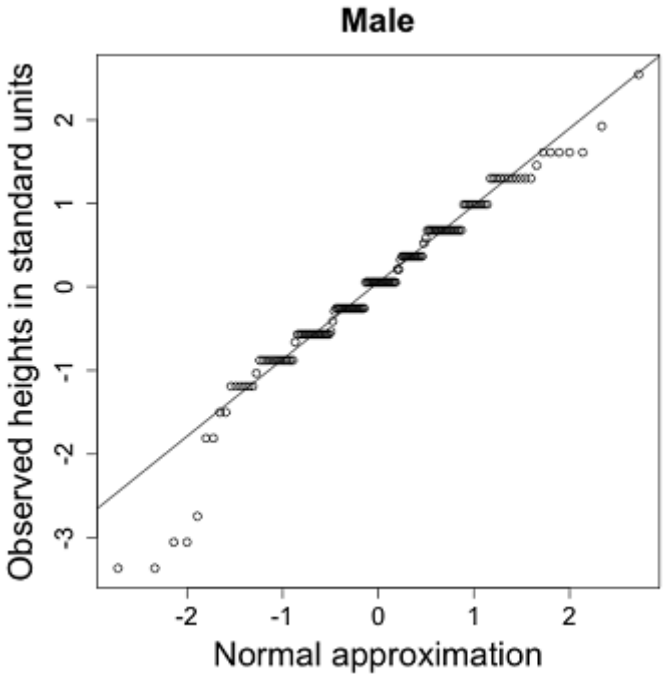

Figure 4:

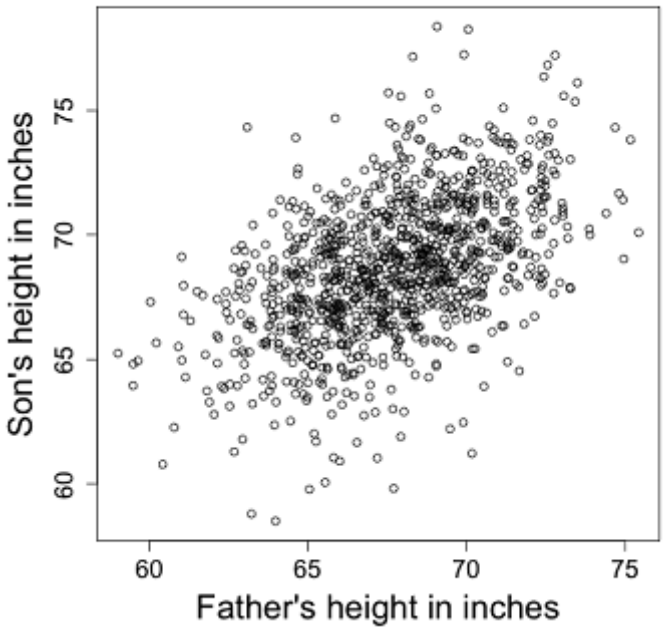

Figure 5:

Two variables

Normal approximation for two variables

Many pairs of data are bivariate normal

- The blue line is the average within each strata
- It is called the regression line

Regression line

The regression line is defined by this formula

$$\frac{Y - \mu_Y}{\sigma_Y} = \rho \frac{X - \mu_X}{\sigma_X}$$

- ρ is called the correlation coefficient
- For fathers and son heights it is 0.5
- For bivariate normal pairs of data these five numbers provide a complete summary:

$$\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$$

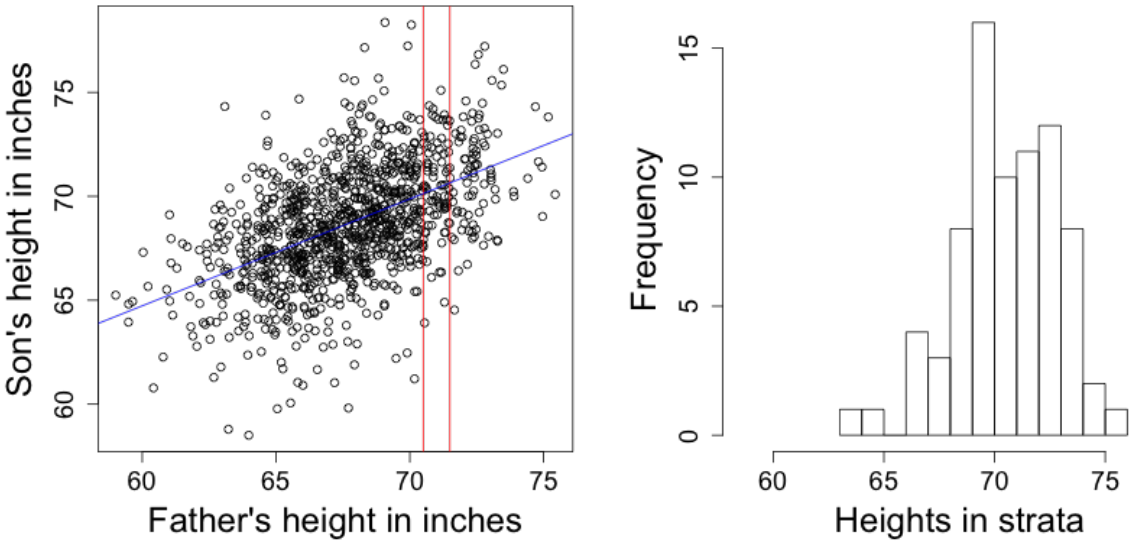

Figure 6:

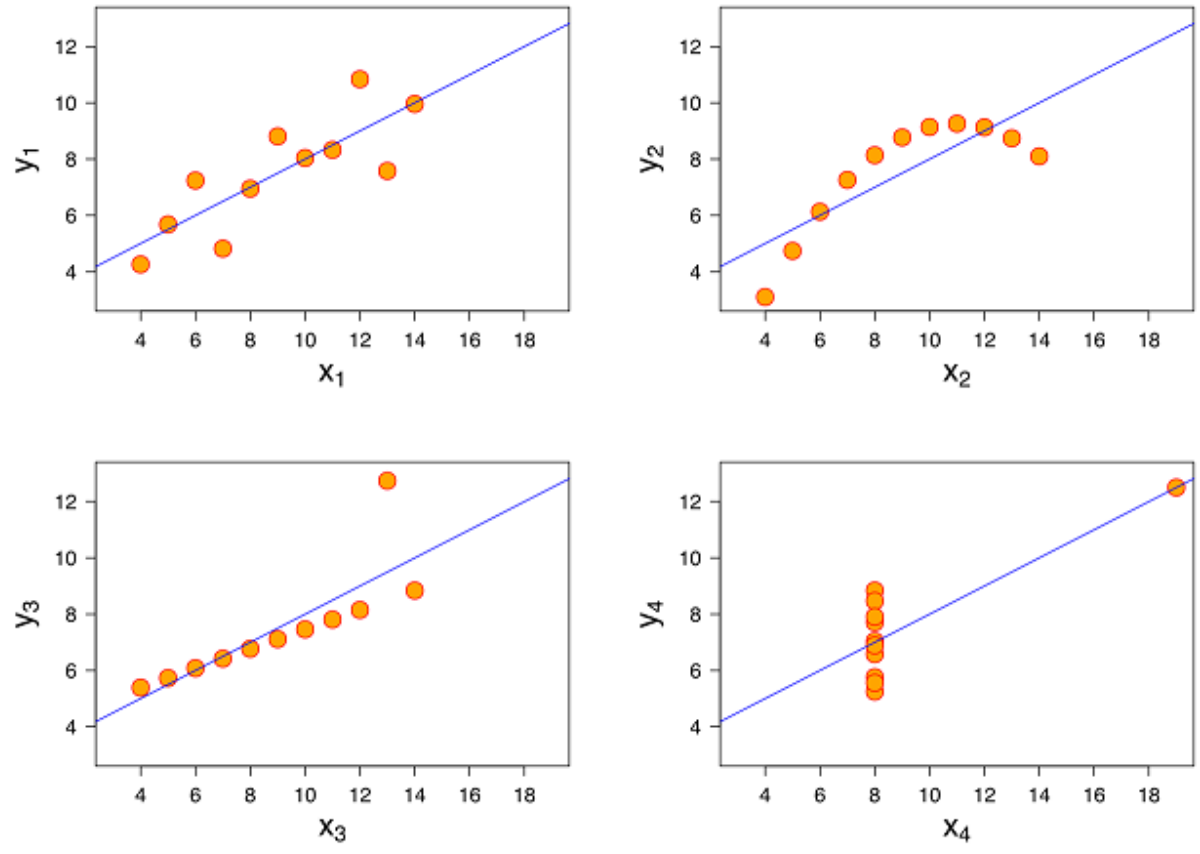

Figure 7:

Anscombe's quartet

Anscombe's quartet

Visualization goals

- Presentation
- Exploratory data analysis

Data

Structured data (or a data set) are exaplined by "variables" and "cases"

- forms a matrix where
  - number of rows = number of cases (individuals)
  - number of columns = number of variables
- Types of variables
  - Nominal (categorical)
  - Ordinal (categorical)
  - Quantitative
- Different types of data are treated differenly when visualized

---

- There are also unstructured types of data
  - text
  - logs
  - images
  - networks
  - trees
  - virtually any object

Example: NHANES

- A survey data collected by the US National Center for Health Statistics (NCHS)
- contained inside the R package `NHANES`
- 10000 individuals, 76 variables

```
varNames <- colnames(NHANES::NHANES)
head(NHANES::NHANES[,c(3,5,9,11,14,17,20)])
```

```
##   Gender AgeDecade    Education    HHIncome HomeRooms Weight Height
## 1   male     30-39  High School 25000-34999         6   87.4  164.7
## 2   male     30-39  High School 25000-34999         6   87.4  164.7
## 3   male     30-39  High School 25000-34999         6   87.4  164.7
## 4   male       0-9         <NA> 20000-24999         9   17.0  105.4
## 5 female     40-49 Some College 35000-44999         5   86.7  168.4
## 6   male       0-9         <NA> 75000-99999         6   29.8  133.1
```

- Type `?NHANES` for information on these demographic variables.
- What are the types of the variables?

Examples of statistical graphics

Let's first browse some options in visualizing the data

Some common graphical elements will be identified, and we will revisit those formally

See DataVis-Supp.pdf

What makes a graphic effective?

A classic text "The Visual Display of Quantitative Information" by Edward Tufte answers this question.

Continue to see DataVis-Supp.pdf

and read the excerpt at http://cs.unm.edu/~pgk/IVCDs14/minitufte.pdf