ST430: Introduction to Regression Analysis, Ch1, Sec 1.1 - 1.6

Luo Xiao

August 19, 2015
Review of Basic Statistics
Definition of *Statistics*

*Statistics* is data science. It involves

- Designing experiments
- Collecting data
- Organizing and classifying data
- Analyzing data
- Interpreting data
- Predicting
Data

Data consists of observed quantities, called *variables*, related to entities called *experimental units*.

Example: Iowa democratic presidential poll with candidates Clinton, Sanders, Biden and etc

- Experimental units ?
- Variables ?
Population and Sample

- Goal: characteristics of a population
- Challenge: observing all experimental units is infeasible
- Ideal solution: a representative sample
- Statistical solution: a random sample
Example

We observe a *sample* from the defined population, and make (*statistical*) *inferences* about the population based on the sample data.

- For example, we might contact 3,000 people by dialing random telephone numbers, reaching 1,000 registered democratic voters.
- We might *infer* that their opinions are representative of the whole state.
- So if the sample shows 50% favors Clinton, 26% favors Sanders, and 24% favors others, we infer that those are the most likely figures in Iowa.
Summarizing qualitative data

A qualitative variable like voting intention is usually summarized as a percentage.

It may be displayed graphically as a bar graph (or histogram), or in a pie chart.
## data
Clinton <- c(favorable = 57, unfavorable = 36, notsure = 7)
## plots
par(mfrow=c(1,2))
barplot(Clinton, main = 'histogram', cex.main=2)
pie(Clinton, main = 'pie chart', cex.main=2)
Summarizing quantitative data

```r
library(fda)
hist(growth$hgtf[31, ], cex.main = 1.25, cex.axis = 1.25, cex.lab = 1.25, main = "Girls' heights (cm) at 18yrs")
```

Girls' heights (cm) at 18yrs
Quantitative data: numerical summaries

The *mean* of data $y_1, y_2, \ldots, y_n$:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$ 

The corresponding population quantity is the *population mean*:

$$\mu = E(Y).$$
The variance of data $y_1, y_2, \ldots, y_n$:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$ 

The corresponding population quantity is the population variance:

$$\sigma^2 = E \left[ (Y - \mu)^2 \right].$$

The standard deviation is the square root of the variance:

$$s = \sqrt{s^2}, \sigma = \sqrt{\sigma^2}.$$
R code

```r
library(fda)
mean(growth$hgtf[31, ]) ## mean

## [1] 166.2944

var(growth$hgtf[31, ]) ## variance

## [1] 39.8145

sd(growth$hgtf[31, ]) ## standard deviation

## [1] 6.309873
```
Interpreting the mean and standard deviation

In *any* data set or population, at least 75% of the data lie within two standard deviations of the mean, by Tchebycheff’s Theorem.

If the data are approximately *normally* distributed, around 95% of the data lie within two standard deviations of the mean.
The standard normal distribution \((\mu = 0, \sigma = 1)\)

```r
x <- seq(-6, 6, by = 0.2)
plot(x, dnorm(x), xlab = "", ylim = c(0, 1), ylab = "", type = "l")
```
Normal distributions

\[ \mu = 0, \sigma = 1 \]
\[ \mu = 3, \sigma = 1 \]
\[ \mu = -4, \sigma = 0.5 \]
Standardizing: if $Y$ has the normal distribution with mean $\mu$ and standard deviation $\sigma$, then

$$Z = \frac{Y - \mu}{\sigma}$$

has mean 0 and standard deviation 1, so it follows the *standard* normal distribution.

One key fact about the standard normal distribution is that

$$P(|Z| \leq 1.96) = .95,$$

and hence also

$$P(\mu - 1.96\sigma \leq Y \leq \mu + 1.96\sigma) = .95$$
For example, if the fuel consumption $Y$ of a car chosen randomly from a year’s production had the normal distribution with $\mu = 37$ mpg and standard deviation $\sigma = 2.4$ mpg, then

$$Z = \frac{Y - 37}{2.4}$$

follows the standard normal distribution.

Then $P(|Z| \leq 1.96) = .95$ implies that

$$0.95 = P\left(\left| \frac{Y - 37}{2.4} \right| \leq 1.96\right)$$

$$= P(37 - 1.96 \times 2.4 \leq Y \leq 37 + 1.96 \times 2.4).$$

In words, there is a 95% chance that the car’s fuel consumption will be between 32.3 and 41.7 mpg.