What is Exploratory Data Analysis?*

Two phases of data analysis:

1. Exploratory Data Analysis (EDA)
   - descriptive statistics (mean, variance)
   - graphical (best visualization of data)
   - data driven

2. Confirmatory Data Analysis (CDA)
   - inferential statistics (CI, hypotheses)
   - theory (and hopefully EDA) driven

The strongest unifying theme underlying exploratory data analysis is expressed in “Look at the data and think about what you are doing.”

*From: Hoaglin, Mosteller, and Tukey, Understanding Robust and Exploratory Data Analysis

Classical Techniques

Like:

- Estimates ($\bar{x}$, $s$)
- $t$-tests and $t$-based CIs

For classical (or “exact”) techniques,

- emphasis on best possible estimate or test when stringent assumptions apply;
- resulting analysis may be garbage when assumptions are violated (sometimes even a little bit).

Robust and Resistant Methods

- aren’t the best possible for given assumptions;
- are the “best” compromise for a wide range of situations;
- often are amazingly close to best possible in most situations.

Resistant to localized misbehaviour (small percentage of gross errors or deviations-from-usual);

Robust even when assumptions (like normality) are flat-out wrong for all the data.

Effect of Recording Error

A sample of 10 observations from $\mathcal{N}(10,4)$:

```r
> x <- round(10*rnorm(10)+4)
> x
[1]  3 17 -12 10 -5 -13  0  8 14 10
> mean(x)
[1] 3.2
> sqrt(var(x))
[1] 10.50714
```

Oops! The number 14 was recorded as 41. (“Bad statistician! Very bad statistician!”)

```r
> y <- x; y[9]
[1] 14
> y[9] <- 41
> mean(y)
[1] 5.9
> sqrt(var(y))
[1] 15.75119
```
Effect of Recording Error

Resistant location estimate (of $\mu$):

```r
> mean(x,trim=.1) # good data
[1] 3.5
> mean(y,trim=.1) # 'bad' data
[1] 3.875
```

Resistant spread estimate (of $\sigma$):

```r
> mad(x) # good data
[1] 10.3782
> mad(y) # 'bad' data
[1] 11.8608
```

EDA and the Three “R”s
(gRaphics, Robustness, and Resistance)

Because EDA deals with data in its rawest form:

- when little is known about it;
- when it hasn’t been “cleaned” up and may be filled with errors or weird behaviour.

the emphasis is on:

- graphical methods (which quickly reveal many problems)
- robust and resistant methods (which work even if problems remain)

From the Textbook


**Chapters 1–4** R (or S or S-PLUS) Language

**Chapters 5–7** Core of the Course

- 5 Univariate Statistics
- 6 Linear Models
- 7 Generalized Linear Models

**Other Topics** One or Two of These:

- 8 Non-Linear and Smooth Regression
- 9 Tree-Based Methods
- 11 Exploratory Multivariate Analysis
- 12 Classification
- 13 Survival Analysis
- 14 Time Series
- 15 Spatial Statistics