INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

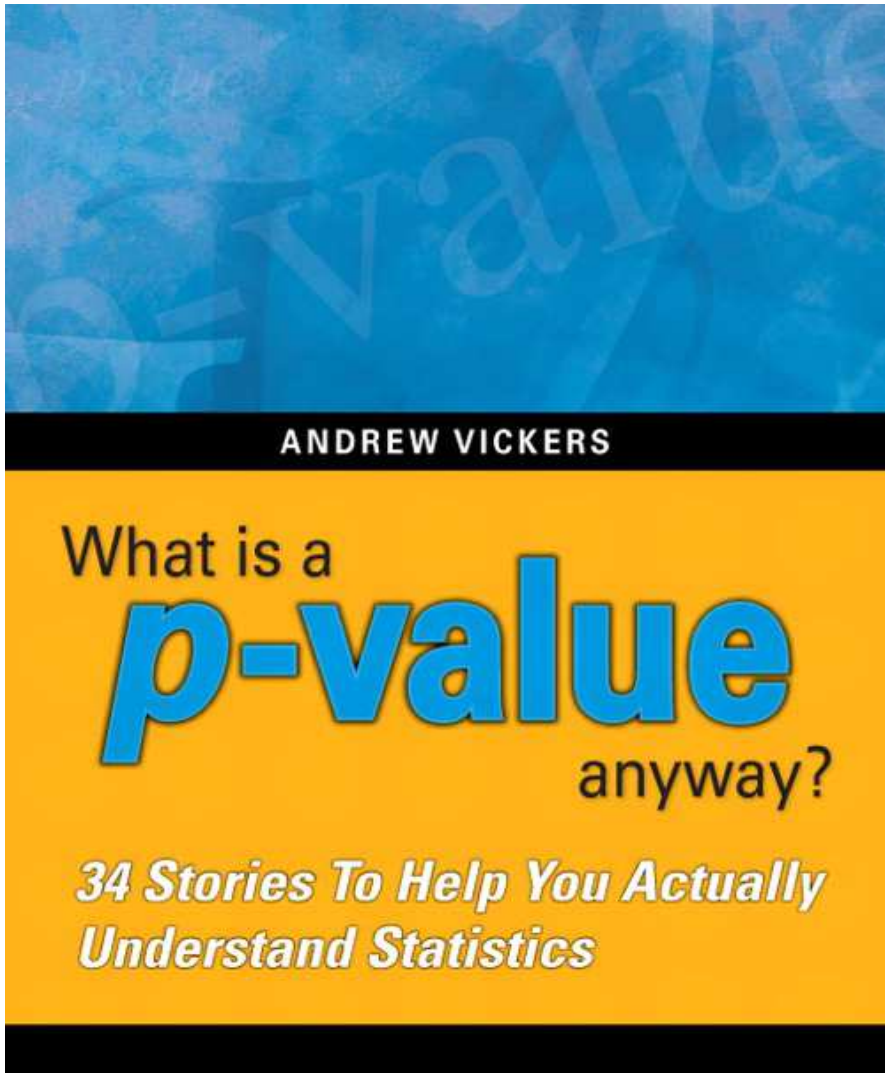
Tento projekt je spolufinancován Evropským sociálním fondem a Státním rozpočtem ČR
InoBio – CZ.1.07/2.2.00/28.0018

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

Data Exploration

Juliette Chamagne
Institute of Evolutionary Biology and Environmental Sciences
University of Zurich

Credits


ANDREW VICKERS
What is a
p-value
anyway?
34 Stories To Help You Actually
Understand Statistics

Statistics

- Estimation vs inference
- Estimation: how big or small something is.
- Inference: making conclusions. Usually with a statistical test or hypothesis.

Data type

- Categorical (e.g. “fertilized” vs “control”)
- Real numbers
  - Continuous (1.34, 147.3…)
  - Discrete (1, 30)

- If real numbers: data restricted? (e.g. positive)

Models

- Dependent variable: the one you want to explain: Y.
- Independent variable: explanatory: X.

- Explain Y: central tendency and measure of dispersion

Central tendency:
About Means and Medians

Central tendency:
About Means and Medians

- Mean (arithmetic mean):

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Median: middle value

Central tendency:
About Means and Medians

$85,000
$50,000
$45,000
$40,000
$35,000
$30,000
$30,000

- Yearly salaries
- **Mean**: $45,000
- Median: $40,000

Central tendency:
About Means and Medians

$1,000,000,000
$85,000
$50,000
$45,000
$40,000
$35,000
$30,000
$30,000

Bill Gates: Outlier

- Yearly salaries
- Mean: $125,039,375
- **Median**: $42,500

Central tendency:
About Means and Medians

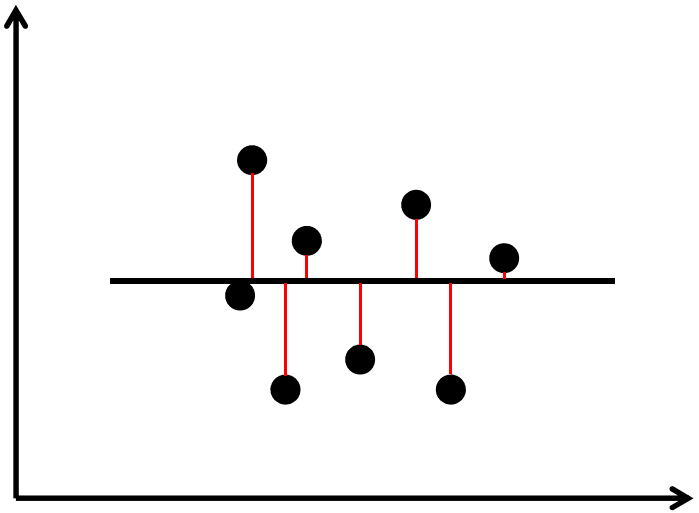
$250,000
$85,000
$50,000
$45,000
$40,000
$35,000
$30,000
$30,000

- Cost of surgery
- **Mean**: $70,625
- Median: $42,500

Measures of dispersion: Standard deviation and interquartile range

- Variance:

$$V = \frac{1}{n} \sum_{i=1}^{p} n_i (x_i - \bar{x})^2$$



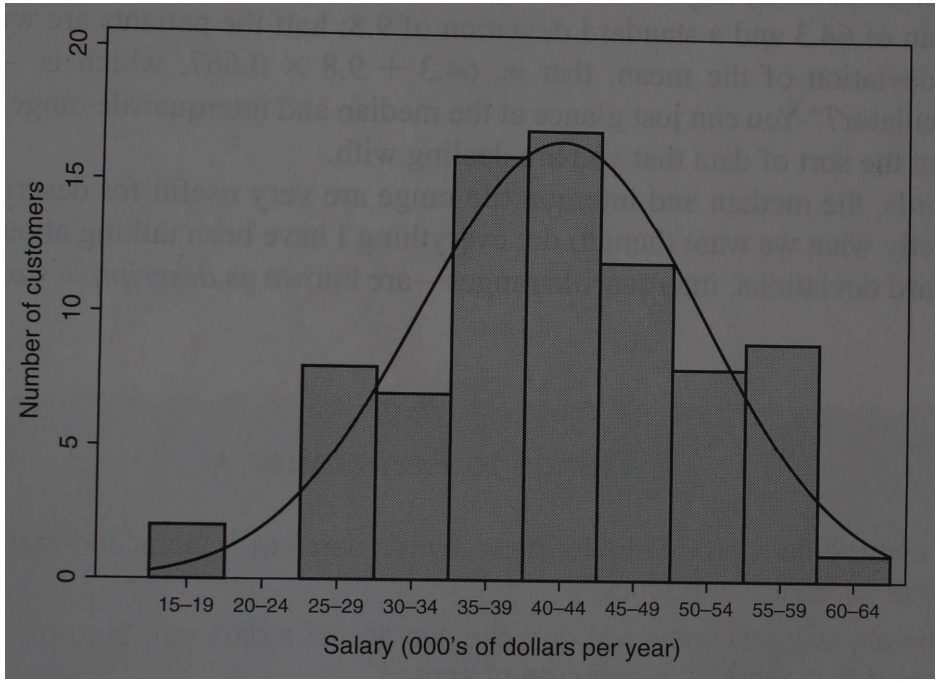
- Standard deviation:

$$SD = \sqrt{V}$$

- Interquartile range: IQR=Q3-Q1
  - Breakdown point of 25%
  - Middle of IQR: median

Measures of dispersion: Standard deviation and interquartile range

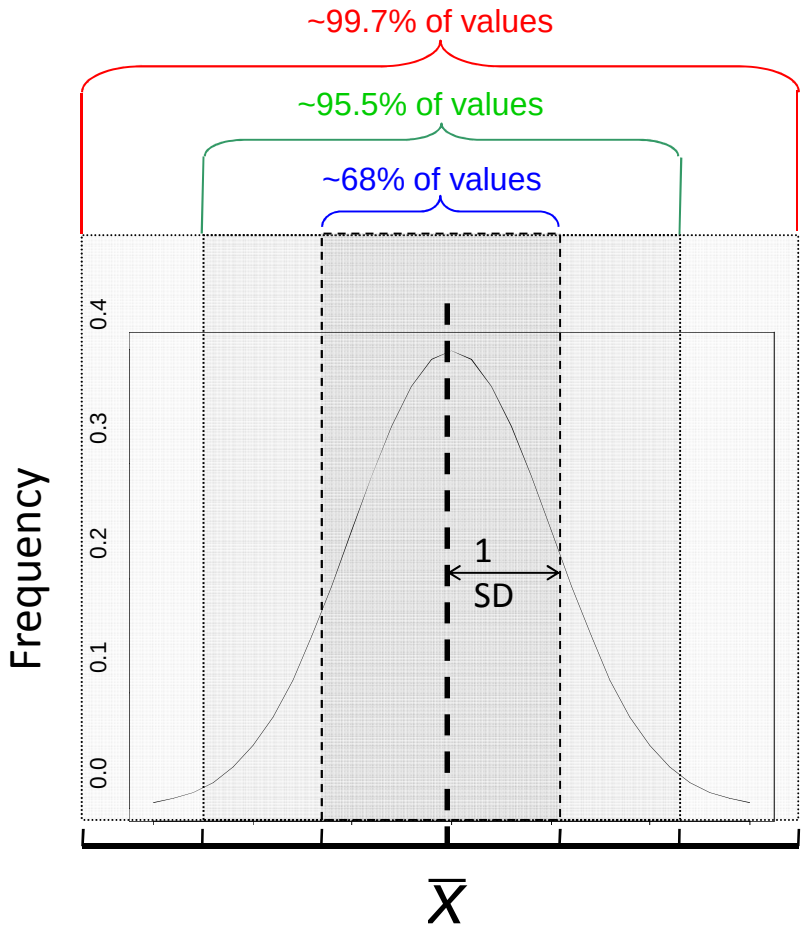
Measures of dispersion: Standard deviation and interquartile range



- Without Bill Gates:
  - Approx. Normal
  - Mean: $42,360
  - SD: $9,616
  - 5% of the salaries below $23,128 or above $61,592

Normal distribution

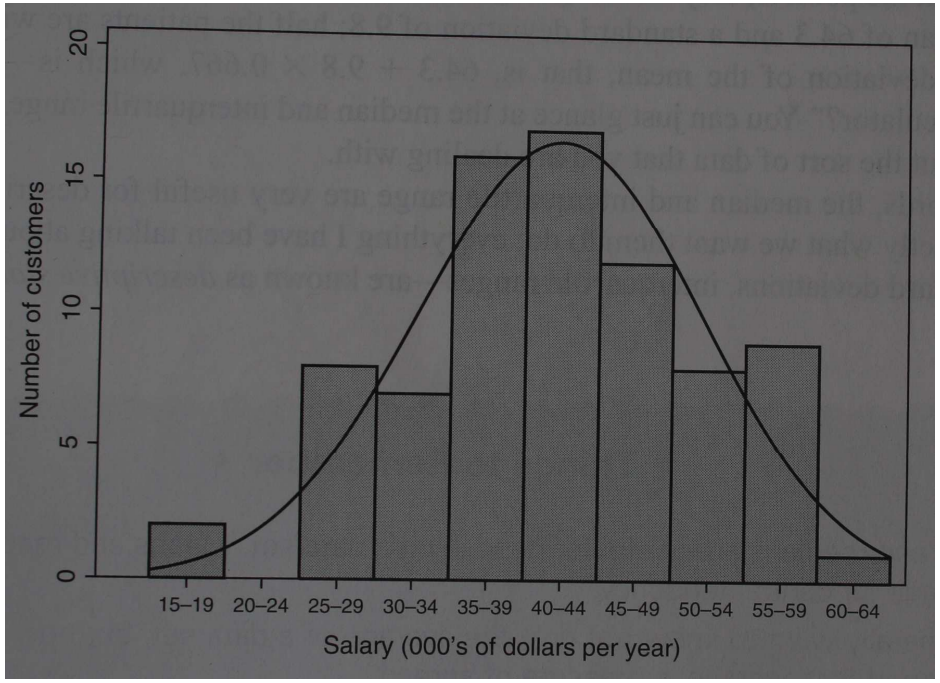
- Sum of lot of random events
- Bell shaped
- Symmetrical
- Mean = median
- 95% of values within ± 2 SD



$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ : Mean σ : SD

Measures of dispersion: Standard deviation and interquartile range



- With Bill Gates:
  - Skewed
  - Mean: $12 million
  - SD: $100 million
  - Can't have a salary of negative $88 million
  - Median: $41,900
  - IQR: $36,000 to $49,300

Correlations

- Measures the strength of association between two continuous variables.
- -1 < correlation between X and Y < 1
- If > 0, positive correlation. Y increases when X increases.
- If < 0, negative correlation. Y decreases when X increases.
- The further away from 0, the stronger the correlation.
- Doesn’t say anything about causality!

Data exploration in R

- c(), list(), data.frame(), matrix()
- read.table(, h=TRUE)
- [], subset()
- ==, %in%
- mean(), range(), dim()
- rep(), seq()
- function(){}

Data exploration in R: Data type

```
> str(dataset)
```

gives the data type of each variable and its values

```
> levels(dataset$categorical.variable)
```

returns the values of a factor

```
> range(dataset$numeric.variable)
```

returns of vector of length 2: min and max

```
> min(dataset$numeric.variable)
> max(dataset$numeric.variable)
```

Data exploration in R: Central Tendency

```
> mean(dataset$numeric.variable)
> mean(dataset$numeric.variable, na.rm=T)
      # na.rm=TRUE: removes NAs
> median(dataset$numeric.variable)


> is.na(dataset$numeric.variable)
      # returns a vector of TRUE and FALSE of the same length
```

Data exploration in R: Measures of dispersion

```
> var(dataset$numeric.variable)
    # variance
> sd(dataset$numeric.variable)
    # standard deviation
> quantile(dataset$numeric.variable)
    # returns min, Q1, median, Q3, and max
```

Data exploration in R: Summary

```
> summary(dataset)
> summary(dataset$variable)
```

returns min, max, quantiles, mean and median

Data exploration in R: Correlation

```
> cor(x,y)
```

returns only the value of the correlation between X and Y.

- X and Y must be two vectors of same length!

```
> cor.test(x,y)
```

returns some information about the significance of the correlation as well. More about that later.

Exercise: Forest

- Fake data.
- 105 forest plots with different number of tree species.
- Productivity estimated as tons of Carbons produced per hectare and per year.
- Forest plots can be divided into 3 categories: “only conifer species”, “only deciduous”, and “mixed conifers and deciduous”.