Data Analysis in R

The Basics of Statistics & Measurement

Ken Stiller

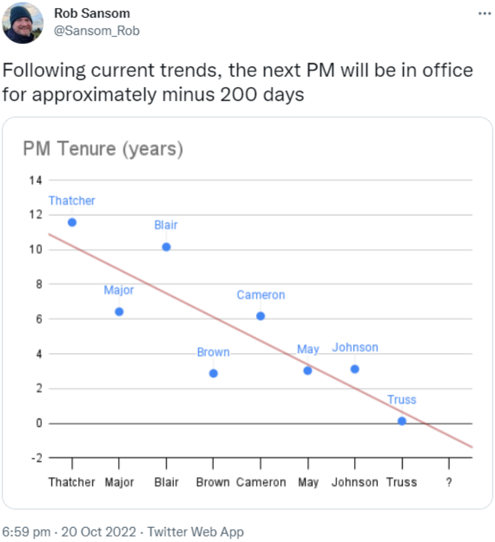
12th January 2024

Syllabus: Data Analysis in R

1. Introduction
2. **Causality & Basics of Statistics**
3. **Measurement**
4. Prediction
5. Multivariate Regression
6. Probability & Uncertainty
7. Hypothesis Testing
8. Assumptions & Limits of OLS
9. Interactions & Non-Linear Effects

Table of Contents

Why Do We Analyse Data?



Definitions

Causality

Refers to the relationship between events where one set of events (the effects) is a direct consequence of another set of events (the causes). (Hidalgo & Sekhon 2012)

Data are Key

The process by which one can use data to make claims about causal relationships. (Hidalgo & Sekhon 2012)

Inferring causal relationships is a central task of science.

Examples

- What is the effect of peace-keeping missions on peace?
- What is the effect of church attendance on social capital?
- What is the effect of minimum wage on employment?

A Counterfactual Logic

Counterfactual Logic

If X had/had not been the case, Y would/would not have happened

Example:*Does college education increase earnings?*

- If high school grads had instead obtained a college degree, how much would their income change?
- If college grads had only obtained a high school diploma, how much would their income change?

A hypothetical example

Imagine two students who are interested in getting a very high score on their thesis. They are considering the courses they should take and they are undecided between *Data Analysis in R* or sticking with *SPSS*.

Y_i : Thesis score is the outcome variable of interest for unit i.

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment (taking Data Analysis in R)} \\ 0 & \text{otherwise.} \end{cases}$$

$$Y_{di} = \begin{cases} Y_{1i} & \text{Potential thesis score for student } i \text{ with Data Analysis in R} \\ Y_{0i} & \text{Potential thesis score for student } i \text{ without Data Analysis in R} \end{cases}$$

Q: What is the effect of taking Data Analysis in R on your thesis score?

Defining the Potential Outcomes

Definition: Treatment

D_i : Indicator of treatment status for unit i

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise.} \end{cases}$$

Defining the Potential Outcomes

Definition: Observed Outcome

Y_i : Observed outcome variable of interest for unit i. (Realized after the treatment has been assigned)

Definition: Potential Outcomes

Y_{0i} and Y_{1i}: Potential Outcomes for unit i

$$Y_{di} = \begin{cases} Y_{1i} & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_{0i} & \text{Potential outcome for unit } i \text{ without treatment} \end{cases}$$

The Fundamental Problem of Causal Inference

The Fundamental Problem of Causal Inference

It is impossible to observe for the same unit i the values $D_i = 1$ and $D_i = 0$ as well as the values Y_{1i} and Y_{0i} and, therefore, it is impossible to observe the effect of D on Y for unit i.

This is why we call this a missing data problem. We cannot observe both potential outcomes, hence we cannot estimate:

$$\tau_i = Y_{1i} - Y_{0i}$$

		Y_{i1}	Y_{i0}
Person 1	Treatment Group ($D = 1$)	Observable as Y	Counterfactual
Person 2	Control Group ($D = 0$)	Counterfactual	Observable as Y

Dealing with this is a core challenge of social science research!

Causal Identification & Internal Validity

- **Association is not causation.**
- *Internal validity* refers to the concern that the difference in outcomes we observe between treated and untreated units are truly caused by the treatment.
- Some threats to internal validity are:
  - Omitted variables
  - Selection bias: Non-random selection into the treatment group
  - Endogeneity and reverse causality
- **Randomised experiments v observational studies**

Statistics: The Basics

Now, we'll briefly discuss the very basics of descriptive statistics:

- Types of variables
- Measures of central tendency
- Quantiles
- Standard Deviation

Types of Variables: Discrete Variables

A **variable** is a measurement of a characteristic of a *unit of analysis* that (usually) varies across unit in a population of units.

There are different levels of measurement:

- **Nominal**: categorical measure, with no ordering
  - e.g . Employed/Unemployed; Single/Married/Divorced
- **Ordinal**: ordered categorical measure
  - The distance between each category is unknown (strongly agree v agree)
  - e.g. many survey questions

Types of Variables: Continuous Variables

- **Interval**: numbers represent a quantitative variable
  - The distance between each level is known and uniform
  - e.g . temperatures, voting cohesion, HDI, measures of democratisation? etc.
  - We can say that it's 10°C more than yesterday
- **Ratio**: There is a meaningful zero mark - which marks complete absence of the measure
  - We can divide measures and express them as multiples
  - e.g. Age: someone might be twice as old as you are whereas this is not the case for temperature (human development?)

Descriptive Statistics

- **Descriptive statistics** are simply that: they describe a large amount of data by summarising it
  - Think of all the values of a variable, which is not very informative - but we somehow want to make sense of them
- Why descriptive stats?
  - Because we're often interested in what a typical unit (e.g .person/country/district etc.) looks like
  - Because it's useful to reduce many measurements to key indicators - either we're interested in them or as a preparatory step
- **Descriptive statistics $\neq$ inferential statistics**

Measures of Central Tendency

- Measuring the *centre* of data - but which one?
  - **Mean:** most common, also referred to as the *average*
    - Sum of measures divided by number of observations

$$\text{mean} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- **Median:**More robust to *outliers*
  - Value at 50% mark of all observed values.

$$\text{median} = \begin{cases} \text{middle value} & \text{if number of entries is odd} \\ \frac{\text{sum of two middle values}}{2} & \text{if number of entries is even} \end{cases}$$

Example: data $= \{0, 1, 2, 3, 100\}$, mean $= 21.2$, median $= 2$

Range & Quantiles

- Measuring the **spread** or **dispersion** of data
  - **Range:** [min(x), max(x)]
  - **Quantile:** 'Portions' of the sorted data: quartile, quantile, percentile, etc.:
    - 25 percentile = lower quartile
    - 50 percentile = median
    - 75 percentile = upper quartile
  - **Interquartile Range (IQR):** Measure of variability and dispersion of the overall variable
    - A dfinition of *outliers*: over 1.5 IQR above upper quartile or below lower quartile

Example:

0%	25%	50%	75%	100%
9.9	16.2	29.2	42.3	75.2

Standard Deviation

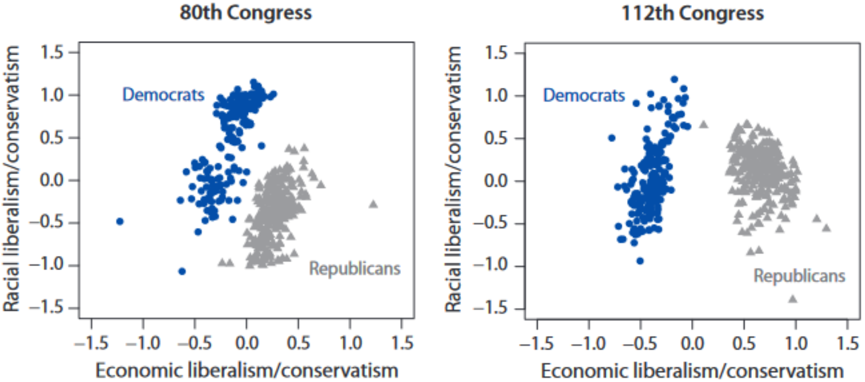
- On average, how far away are data points from their mean?

$$\text{standard deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Root-Mean-Square (RMS) of deviation from average
- Sometimes it's divided by n instead of $n - 1$
- Variance = standard deviation^2

Doing Research is a process.
What role does measurement play?

Example: Measuring Ideology



Source: Imai, p.99

Visualizing Univariate Distributions

- Descriptive statistics are useful, but sometimes more helpful to visualize the distribution of a variable.
- There are several ways to do this as you have learnt:
  - Barplots
  - Histograms
  - Boxplots[...]
- We'll use survey data from Afghanistan as an example

Barplot

- Visualize the distribution of a categorical (*factor*) variable
    - In this case, whether respondent reported victimization by the coalition of international troops (ISAF)

```
barplot(prop.table(table(ISAF = afghan$violent.exp.ISAF,
                         exclude = NULL)),
        names.arg = c("No harm", "Harm", "Non-response"),
        main = "Civilian victimization by the ISAF",
        xlab = "Response category",
        ylab = "Proportion of the respondents",
        ylim = c(0, 0.7))
```

Barplot II


Civilian victimization by the ISAF
Proportion of the respondents
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
No harm
Harm
Non−response
Response category

Histogram

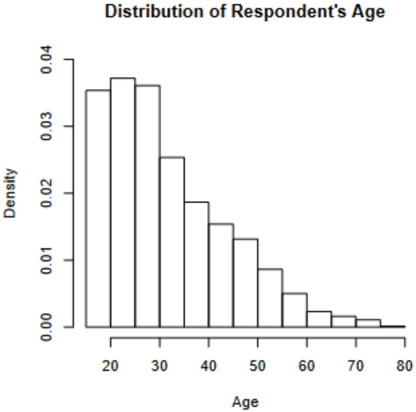
- Visualize the distribution of a continuous variable
- It might help to think about how to create a histogram by hand:
  1. create bins across the variable of interest
  2. count number of observations in each bin
  3. frequency = bin height

$$\text{density} = \frac{\text{proportion of observations in bin}}{\text{bin width}}$$

- In R, we use `hist()` with `freq= FALSE`

```
hist(afghan$age, freq = FALSE, ylim = c(0, 0.04),
     xlab = "Age", main =
"Distribution of Respondent's Age")
```

Histogram II

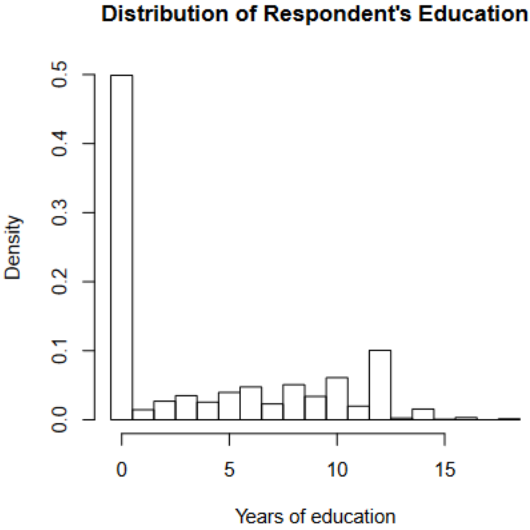


Let's Be Clear About Density

- The areas of the blocks sum to 1 or 100%
- Density ≠ Percentage
- The height of the blocks equals the percentage divided by the bin width: in this case, "percent per year"
- More generally, *percentage per horizontal unit*
- We can also choose the bin locations on our own via the `breaks` (locations of bin breaks) or `nclass` (number of bins) arguments

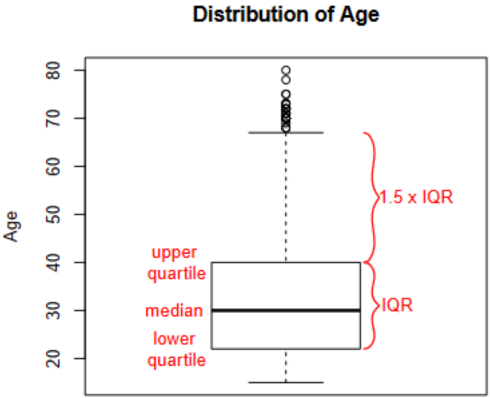
```
hist(afghan$educ.years, freq = FALSE,
     breaks = seq(from = -0.5, to = 18.5, by = 1),
     xlab = "Years of education",
     main = "Distribution of Respondent's Education")
```

Density II


Distribution of Respondent's Education
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
Years of education

Boxplot

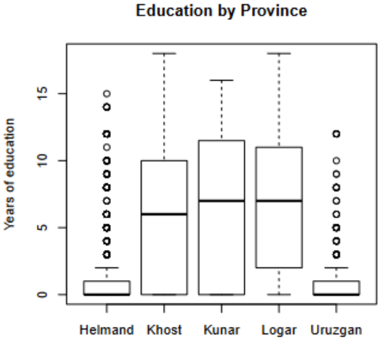
- Characterises the distributions of continuous variables at
- Features:
  - box, whiskers, outliers



Boxplot II

- *Boxplots* also can give you a good overview by groups
- Useful for comparison across multiple categories: `boxplot(y ~ x, data = d)`

```
boxplot(educ.years ~ province, data = afghan,
        main = "Education by Province",
        ylab = "Years of education")
```



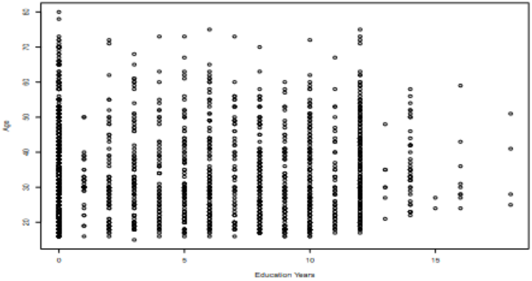
Bivariate Relationships

- More than in univariate distributions, we are often interested in how *two variables* *relate* to one another
- There, again, are various ways to do this, some of which are:
  - Scatterplots
  - Correlation coefficients
- We'll continue to use the Afghanistan survey data as an example

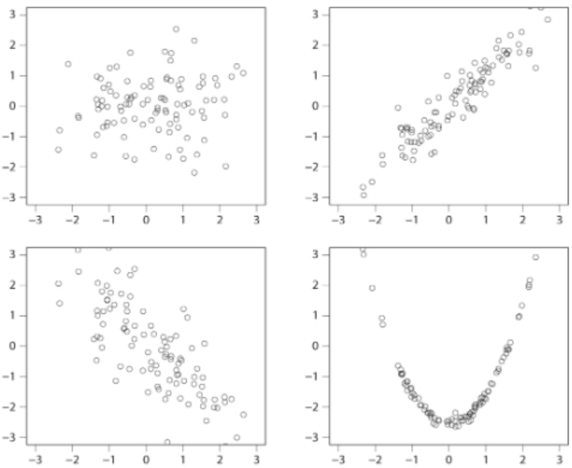
Scatterplot

- Direct graphical comparison of two variables, for same units
- Can simply use `plot()` function

```
plot(afghan$educ.years, afghan$age,
     xlab = "Education Years", ylab = "Age")
```



Scatterplot II



Correlation

- On average, how do two variables move together?
- Mathematical definition of the correlation coefficient:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \text{ mean of } x}{\text{standard deviation of } x} \times \frac{y_i - \text{ mean of } y}{\text{standard deviation of } y}\right)$$

$=$ mean of products of z-scores

- As with standard deviation, sometimes $n-1$ is replaced with n

Correlation II

- On average, how do two variables move together?
- Positive correlation: When x is larger than its mean, y is likely to be larger than its mean
- Negative correlation: When x is larger than its mean, y is unlikely to be larger than its mean
- Positive [negative] correlation: data cloud slopes up [down]
- High correlation: data cluster tightly around a line

Example: Correlation of Age and Education

- Compute the correlation in R:

```
cor(afghan$educ.years, afghan$age,
    use = "pairwise")
## [1] 0.04569074
```

- Low correlation! What is low/high?

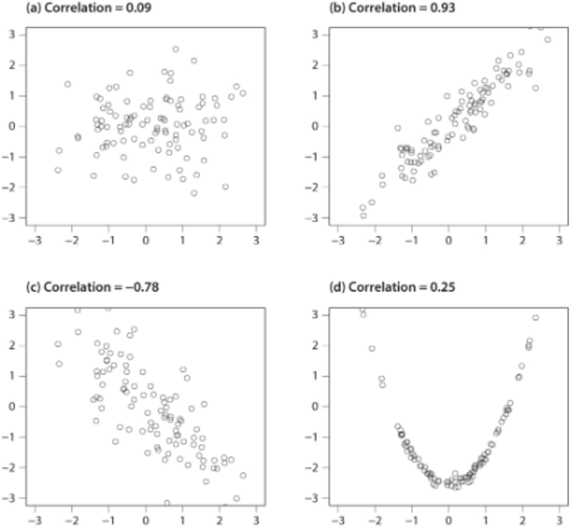
Properties of the Correlation Coefficient

- Correlation is - by design - between -1 and 1
- Order does not matter: $\mathrm{cor(x, y) = cor(y, x)}$
- Not affected by changes of scale:

$$\mathrm{cor}(x, y) = \mathrm{cor}(ax + b, cy + d)$$

  for any numbers `a`, `b`, `c` and **d**
- Measures don't matter (but ideally do): C v F, cm v inch, *e* v $
- Keep in mind: Correlation measures *linear* association!

Correlation III



ggplot2

Note: `ggplot` is an ubiquitous package for creating figures in `R` that is more powerful and versatile than base `R` - you'll find some examples on the course page.