

Data Analysis in R

Sampling & Visualisation

Ken Stiller

16th January 2026

Syllabus: Data Analysis in R

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

Recap Causality

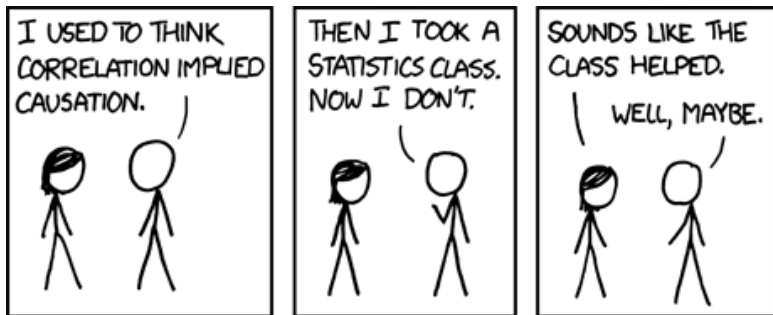


Table of Contents

Introduction

Sampling

Univariate Distributions

Bivariate Relationships

Measurement

Wrap Up

Samples & Populations

- ▶ Oftentimes, when numbers are reported they relate to a **sample** of a population
 - ▶ **Cost:** For elections polls, we could ask all eligible voters
 - ▶ The last UK census cost roughly 480 million
 - ▶ The 2017 UK election cost more than 140 million
 - ▶ **Speed:** It took about 5 years to process all the census data
- ▶ Therefore, we use samples to make **inferences** about a population
- ▶ Sample statistics are estimates of population parameters
 - ▶ For a statistic to be useful, the sample needs to be **representative** of the population

'Bad' Example: The 1936 Literary Digest Poll

- ▶ Had correctly predicted every US presidential election winner since 1916
- ▶ Over 10 million questionnaires sent to subscribers
- ▶ Final sample size: over 2.3 million returned.
- ▶ Predicted landslide for Landon versus Roosevelt

	FDR's vote share
Literary Digest	43
George Gallup	56
Actual Outcome	62

Basic Sampling

- ▶ **Simple random sampling**
 - ▶ Every unit has an *equal* selection probability
- ▶ **Probability sampling** to ensure representation
 - ▶ Every unit in the population has a *known non-zero probability* of being selected

Sampling & Biases

Population	Sample	Potential Bias
Target population		
↓		Frame bias
Frame population →	Sample	Sampling bias
	↓	Unit non-response
	Respondents	
	↓	Item non-response
	Completed items	
		Response bias

Solutions to SRS Problems

- ▶ **Stratified** random sampling
 - ▶ Two stages: classify population into groups, then select by SRS *within* groups
 - ▶ e.g. 'over-sample' cat owners: divide population by cat ownership, use SRS to select 50 cat-owning and 50 cat-free households
- ▶ **Cluster** random sampling
 - ▶ If population members are naturally *clustered*, we can SRS clusters, and then SRS respondents within selected clusters
 - ▶ e.g. pupils are naturally grouped by school, so randomly select 5 schools, then randomly pick 10 children from each school

Non-Response Bias

- ▶ **Unit** non-response
 - ▶ Certain members of chosen sample may not respond to survey at all
 - ▶ Can create bias even if sample is representative
 - ▶ This was a problem with the Literary Digest survey (alongside sampling bias)
- ▶ **Item** non-response
 - ▶ Respondents may choose not to answer certain questions
 - ▶ Can create bias even if sample is representative
 - ▶ Sensitive questions \rightsquigarrow non-response
- ▶ If those who refuse to answer are **systematically different** from those who answer, resulting inference is (very likely) biased

Example: Item Non-Response

- ▶ Civilian Survey in Afghanistan (Lyll, Blair, Imai 2013)
- ▶ Can the hearts and minds of civilians be won?
- ▶ One question on victimization by Taliban and ISAF
 - ▶ More violent areas \rightsquigarrow larger share of non-responses

```
afghan <- read.csv("data/afghan.csv")
tapply(afghan$violent.exp.taliban, afghan$province,
       mean, na.rm = TRUE)

## Helmand  Khost  Kunar  Logar Uruzgan
##  0.5042  0.2332  0.3030  0.0802  0.4545

tapply(is.na(afghan$violent.exp.taliban),
       afghan$province,
       mean)

## Helmand  Khost  Kunar  Logar Uruzgan
##  0.03041  0.00635  0.00000  0.00000  0.06202
```

Response Bias

- ▶ Even when respondents do respond, need to be wary of **misreporting**
- ▶ Responses can be affected by many things, including:
 - ▶ Question ordering
 - ▶ Interview/Sampling setting
 - ▶ Identity of interviewer, etc.
- ▶ Sensitive questions \rightsquigarrow **social desirability bias**
 - ▶ e.g. Racial prejudice, corruption, even turnout, income, wealth
 - ▶ e.g. "Shy Trump voter"

Visualizing Univariate Distributions

- ▶ Descriptive statistics are useful, but sometimes it's 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 the Afghanistan survey data 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



Histogram

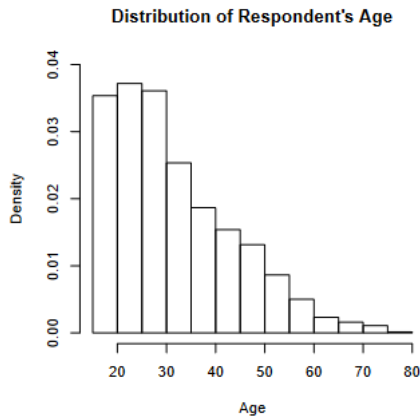
- Visualize the distribution of a **continuous** variable
- It might help think about how to create a histogram by hand:
 1. create bins across the variable of interest
 2. count number of observations in each bin
 3. **frequency** = bin height

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

- 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

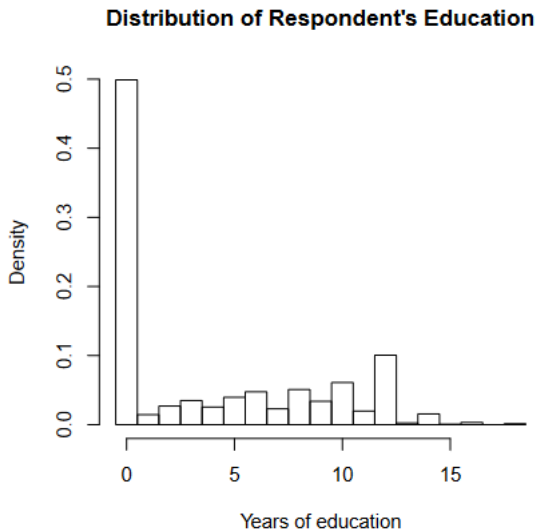


Let's Be Clear About Density

- ▶ The areas of the blocks sum to 1 or 100%
- ▶ Density \neq 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

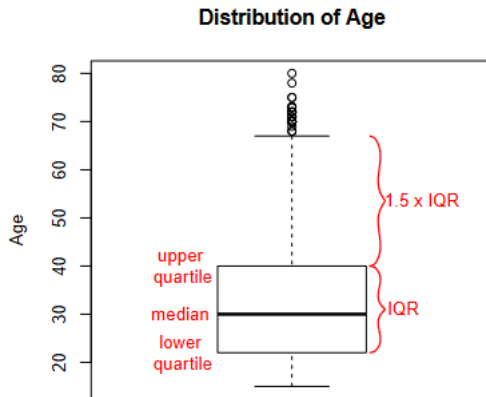
```
hist(afghan$educ.years, freq = FALSE,  
     breaks = seq(from = -0.5, to = 18.5, by = 1),  
     xlab = "Years of education",  
     main = "Distribution of Respondent's Education")
```

Density II



Boxplot

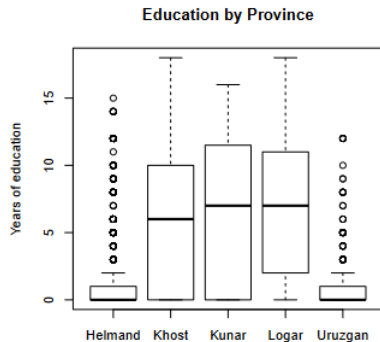
- ▶ Characterises the distributions of continuous variables at
- ▶ Features:
 - ▶ box, whiskers, outliers



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")
```



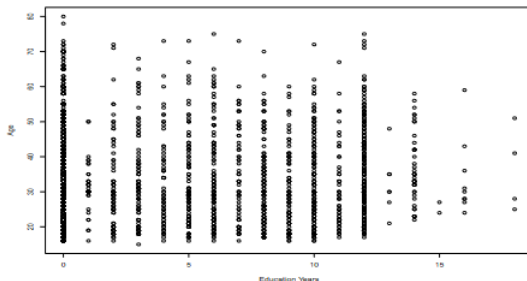
Bivariate Relationships

- ▶ 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

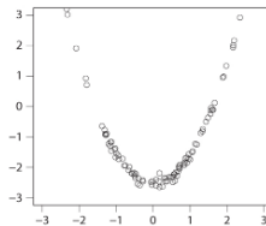
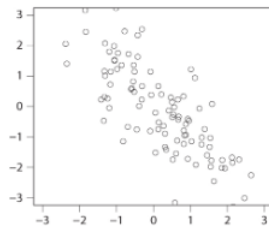
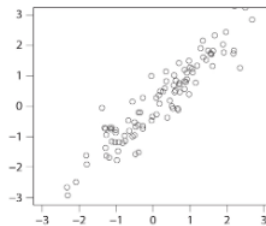
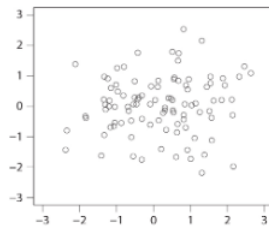
Scatterplot

- ▶ 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: $\text{cor}(x, y) = \text{cor}(y, x)$
- ▶ Not affected by changes of scale:

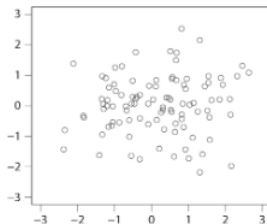
$$\text{cor}(x, y) = \text{cor}(ax + b, cy + d)$$

for any numbers **a**, **b**, **c** and **d**

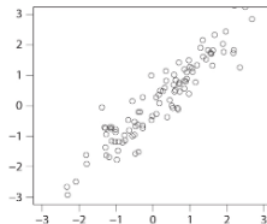
- ▶ Units don't matter (but we care): $^{\circ}\text{C}$ v $^{\circ}\text{F}$, cm v inch, € v $\text{\$}$
- ▶ **Keep in mind:** Correlation measures *linear* association!
- ▶ It does not reveal the magnitude of association either!

Correlation III

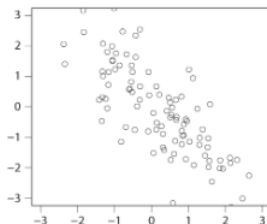
(a) Correlation = 0.09



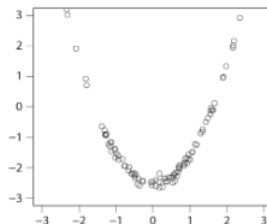
(b) Correlation = 0.93



(c) Correlation = -0.78

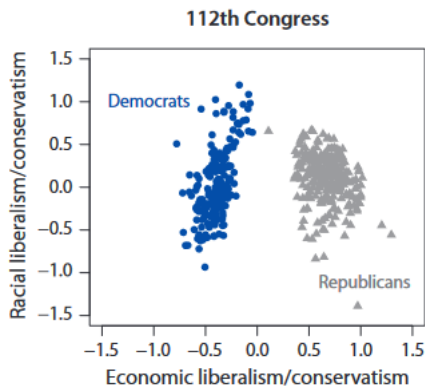
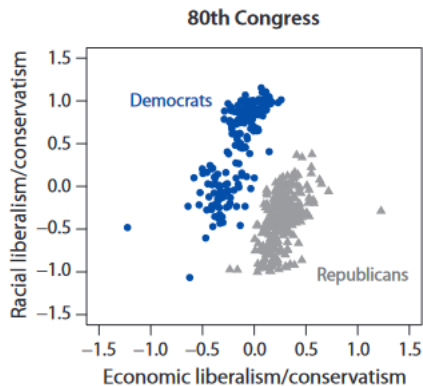


(d) Correlation = 0.25



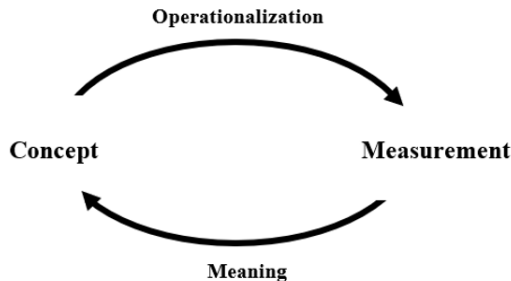
**Doing research is a process.
What role does measurement play?**

Example: Measuring Ideology

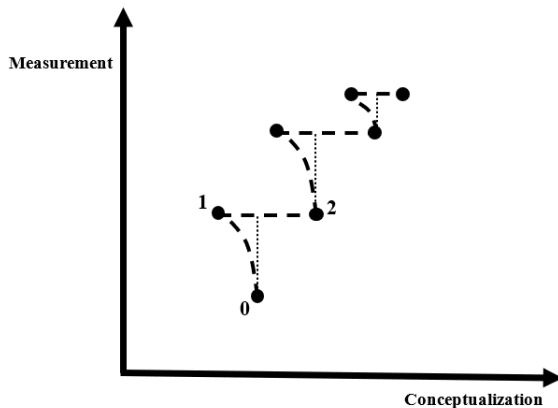


Source: Imai, p.99

Data are not separate from Concepts



How Do Empirics Matter?



Thinking About Concepts + Measurement

Do conceptualisation and even measurement affect how we conceive of things? This is a question you should always have in mind when making decisions on either.

Even seemingly banal concepts can be given a distinct perspective/meaning by the way we measure them: [Example](#)

Wrap Up

- ▶ Key points:
 - ▶ Sampling is necessary; probability sampling the gold standard
 - ▶ Multiple sources of bias, even with SRS
 - ▶ Visual descriptions and summaries of variables
 - ▶ Correlations as useful ways to describe relationships between variables
- ▶ Next, we'll be talking about:
 - ▶ Bivariate linear regressions

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.