

Models of Reality

From Data to Science

Prof. Dr. Jochen Kruppa-Scheetz

2025-12-05

Table of contents

I	From science to data by models	5
	References	6
1	What is science?	7
1.1	General background	8
1.2	Theoretical background	9
1.3	R packages used	9
1.4	Data	9
1.5	Alternatives	9
1.6	Glossary	9
1.7	The meaning of “Models of Reality” in this chapter.	9
1.8	Summary	9
	References	9
2	What is data?	10
2.1	General background	10
2.2	Theoretical background	10
2.3	R packages used	10
2.4	Data	10
2.5	Alternatives	11
2.6	Glossary	11
2.7	The meaning of “Models of Reality” in this chapter.	11
2.8	Summary	13
	References	13
3	What is a model?	14
3.1	General background	14
3.2	Theoretical background	17
3.3	R packages used	21
3.4	Data	21
3.5	Alternatives	21
3.6	Glossary	21
3.7	The meaning of “Models of Reality” in this chapter.	21
3.8	Summary	21
	References	21

II	Tales of data	22
	References	23
4	Body weight of fleas	24
4.1	General background	24
4.2	Theoretical background	24
4.3	R packages used	24
4.4	Data	24
4.4.1	Linear	24
4.4.2	Non-linear	24
4.5	Alternatives	26
4.6	Glossary	26
4.7	The meaning of “Models of Reality” in this chapter.	26
4.8	Summary	26
	References	26
5	Fleas of animals	27
5.1	General background	27
5.2	Theoretical background	27
5.3	R packages used	28
5.4	Data	28
5.4.1	Jump performances of fleas	28
5.4.2	Measurements on animal fleas	31
5.4.3	Fleas in urban and rural habitats	32
5.5	Data availability	33
5.6	Alternatives	34
5.6.1	Why linear is a bad idea?	34
5.6.2	Why is a <code>tibble()</code> bad idea?	34
5.7	Glossary	34
5.8	The meaning of “Models of Reality” in this chapter.	34
5.9	Summary	34
	References	34
6	Eating habits of fleas	35
6.1	General background	35
6.2	Theoretical background	36
6.3	R packages used	36
6.4	Data	36
6.4.1	Linear	36
6.4.2	Non linear	36
6.4.3	International study	37
6.5	Alternatives	40
6.6	Glossary	41

6.7	The meaning of “Models of Reality” in this chapter.	41
6.8	Summary	41
	References	41
III	Speaking to data	42
7	Programing in the 21st century	44
7.1	General background	44
7.2	Theoretical background	44
7.3	R packages used	44
7.4	Data	44
7.5	Alternatives	44
7.6	Glossary	44
7.7	The meaning of “Models of Reality” in this chapter.	45
7.8	Summary	45
	References	45
IV	Hypothese testing	46
8	Statistical testing	48
8.1	General background	48
8.2	Theoretical background	52
8.2.1	Fisher’s approach: The ‘significance test’	52
8.2.2	Neyman-Pearson’s approach: The ‘hypothesis test’	52
8.2.3	Today’s ‘hybrid chaos’	54
8.3	R packages used	54
8.4	Data	54
8.5	Alternatives	54
8.6	Glossary	55
8.7	The meaning of “Models of Reality” in this chapter.	55
8.8	Summary	55
	References	55
V	Template Preface	56
9	Template chapter	58
9.1	General background	58
9.2	Theoretical background	58
9.3	R packages used	58
9.4	Data	58
9.5	Alternatives	58

9.6	Glossary	58
9.7	The meaning of “Models of Reality” in this chapter.	58
9.8	Summary	59
	References	59

Appendices 60

A Why does it look like this? 60

A.1	Used R packages	60
A.2	Used theme of the visualisations	60
A.3	Used color palettes	60
	References	61

Welcome

“Life is difficult.” — M. Scott Peck’s, The Road Less Traveled

This will be a long journey... it’s a hell of a time.

Preface

It's too early to write a preface. I started writing this book a few weeks ago. From my perspective, it will be a good one, but it will take some time. In the meantime, you are welcome to see what's happening here and what I'm writing. Sometimes things will be in the wrong place. This is a normal part of the writing process. Just so you know, I started writing on 15 November 2025.

text¹

“The first principle is that you must not fool yourself — and you are the easiest person to fool. So you have to be very careful about that.” —²

“Hic sunt dracones!” — *Hic Sunt Dracones on the Hunt–Lenox Globe* (eng. “Here be dragons”)

3

References

Part I

From science to data by models

Last modified on 22. November 2025 at 19:08:33

“I’d like to solve the puzzle.” — Wheel of Fortune

In 2008, Cadiergues and his team won the Ig Nobel Prize at Harvard University in Cambridge for their work with dog and cat fleas⁴. This book tells the story of what had happened if they won the Nobel Prize in Stockholm instead. The Nobel Prize sparked extensive research in many scientific fields. Fleas became all the rage. Lots of data was collected, fleas were measured, and questions were asked that no one had ever asked before. Let’s check out this data and see what it tells us about fleas. Come along to Adventure Land and get ready to be amazed.

None of this happened in our reality. At least not in our branch of reality, if you believe Everett⁵. But since this book is also about models of reality, we can question the whole of reality and create a new one that fits.

References

1 What is science?

[conflicted] Will prefer dplyr::filter over any other package.

Last modified on 27. November 2025 at 15:18:05

“Reality is negotiable.” — Tim Ferriss

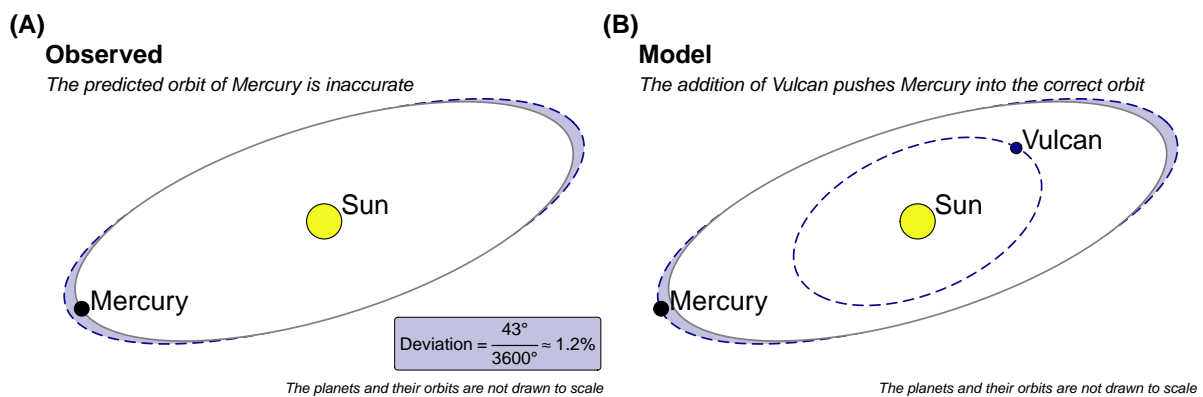


Figure 1.1: foo (A) foo (B) foo

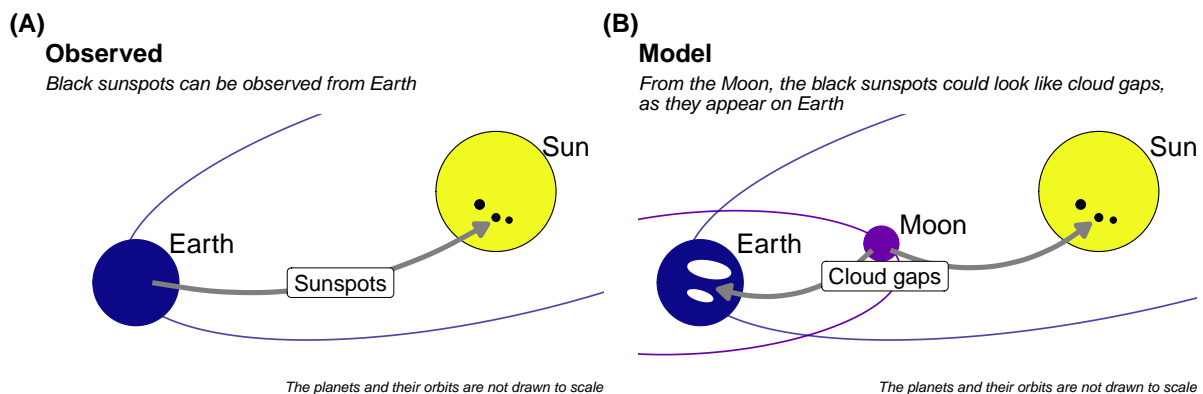


Figure 1.2: foo (A) foo (B) foo

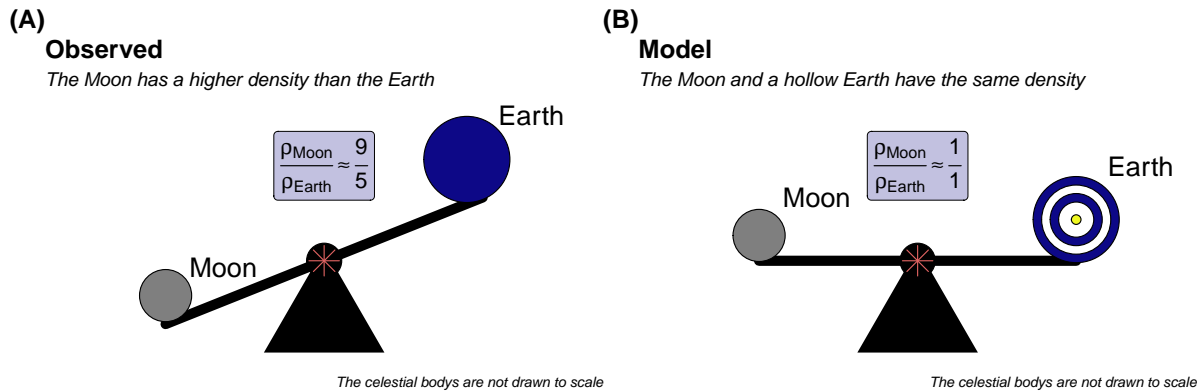


Figure 1.3: foo (A) foo (B) foo

1.1 General background

“Any sufficiently advanced technology is indistinguishable from magic.” —⁶, Clarke’s third law

Full quote

“Reality is negotiable. Outside of science and law, all rules can be bent or broken, and it doesn’t require being unethical.” — Tim Ferriss

Idea of hypotheses

Science is guessing and falsification

⁷ [What is this thing called Science?](#)

⁸ [What is science](#)

⁹ [What is science?](#)

¹⁰ [Models Demystified: A Practical Guide from Linear Regression to Deep Learning](#)

¹¹ [Statistical Thinking for the 21st Century](#)

¹² [The beginning of infinity: Explanations that transform the world](#)

[David Deutsch > Quotes](#)

1.2 Theoretical background

1.3 R packages used

1.4 Data

1.5 Alternatives

Further tutorials and R packages on XXX

1.6 Glossary

term what does it mean.

1.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

1.8 Summary

References

2 What is data?

[conflicted] Will prefer `dplyr::filter` over any other package.

Last modified on 18. November 2025 at 07:21:12

“The limits of my language mean the limits of my world.” — Ludwig Wittgenstein

2.1 General background

2.2 Theoretical background

2.3 R packages used

2.4 Data

```
jump_weight_tbl <- tibble(x = c(0.6, 1, 2.3, 3.5, 5.2, 7.1, 8.4, 9.2, 10),  
                          y = 0.15*x^3 - 2.2*x^2 + 8.8*x + 3.2 + rnorm(9, 0, 0.5)) |>  
  mutate_all(round, 1) |>  
  rename(weight_mg = x, jump_length_cm = y)
```

Table 2.1: foo.

weight_mg	jump_length_cm
0.6	8.5
1.0	9.6
2.3	13.7
3.5	13.2
5.2	11.0
7.1	8.3
8.4	10.9

Table 2.1: foo.

weight_mg	jump_length_cm
9.2	14.3
10.0	21.1

Equation 2.1

$$y = 0.15 \cdot x^3 - 2.2 \cdot x^2 + 8.8 \cdot x + 3.2 \quad (2.1)$$

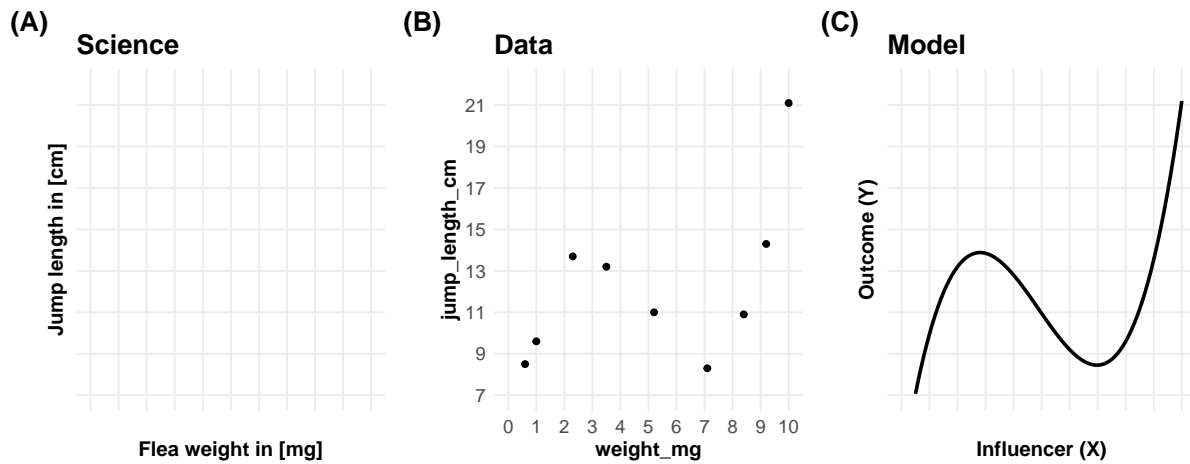


Figure 2.1: foo.

2.5 Alternatives

Further tutorials and R packages on XXX

2.6 Glossary

term what does it mean.

2.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

Science with data and model

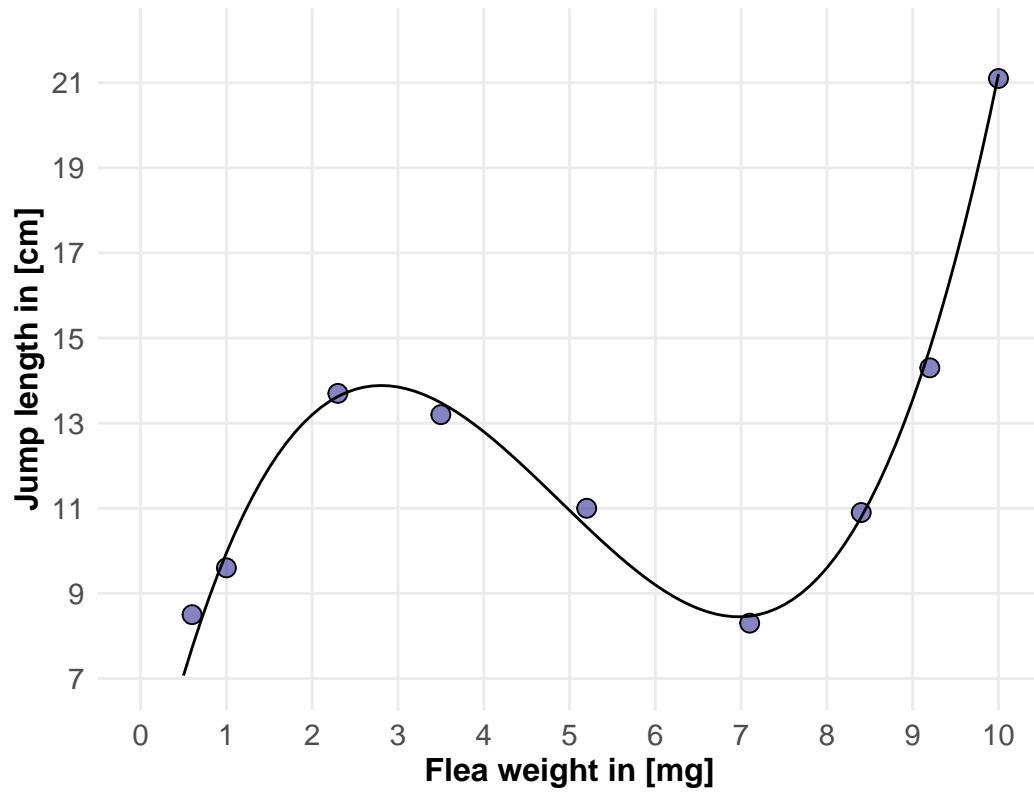


Figure 2.2: foo.

2.8 Summary

References

3 What is a model?

[conflicted] Will prefer `dplyr::filter` over any other package.

Last modified on 26. November 2025 at 13:17:46

“A quote.” — Dan Meyer

Imagine you are standing on a globe but still believe it is flat. You are standing on solid ground, with boiling rocks a few kilometers below you. A few kilometers above, you cannot breathe anymore. You are not moving, but the Earth is moving at speeds of up to 828,000 km/h, depending on the reference point. Your reality is a serious misinterpretation.

3.1 General background

Table 3.1: Table of terms used in the statistical modelling. The terms in bold are used here. Depending on the scientific background, the usage of these terms can vary widely.

Symbol	Name Application	Description
y	out- come, re- sponse, end- point, de- pen- dent vari- able	<i>The right-hand side (abbr. RHS) of the model. Describing the values measured in an experiment or study.</i>

Symbol	Name	Application	Description
x	in-fluencer , in-fluential variable, risk factor, fixed effect, independent variable		<i>The left-hand side (abbr. LHS) of the model. Describing the influential variables in an experiment or study.</i>
z	random effect		<i>A factor that provides a description of an grouping variable, which is not part of the controlled experimental setting.</i>
x	explanator , explanatory variable	Explanation	<i>The influencer is used to describe or explain the outcome.</i>
x	predictor , predictive variable	Prediction	<i>The influencer is used to predict the outcome.</i>

Symbol	Name	Application	Description
x	fo- cal ex- plana- tor, fo- cal pre- dic- tor, fo- cal vari- able	Main effect	<i>In a model with multiple influencers, the focal variable is the variable of primary interest.</i>
c	co- vari- ate, co- vari- able	Continuous x	<i>The influencer is a numeric variable with continuous values.</i>
f_A	fac- tor A, fac- to- rial vari- able, cat- e- gori- cal vari- able	Categorical x	<i>The influencer is discrete, functioning as a grouping variable, such as an experimental group or a treatment.</i>
$A.1$ to $A.j$	lev- els, groups, treat- ment groups	Factor f_A	<i>The discrete groups included in one factor A.</i>

A sentence why we use y and not x for mean and other stuff.

¹³ [What is a statistical model?](#)

¹⁴ [What do we mean by a statistical model?](#)

¹⁵ [What Is the Purpose of Statistical Modeling?](#)

¹⁶ [Where do statistical models come from? Revisiting the problem of specification](#)

¹⁷ [Statistical modelling](#)

3.2 Theoretical background

Model notation

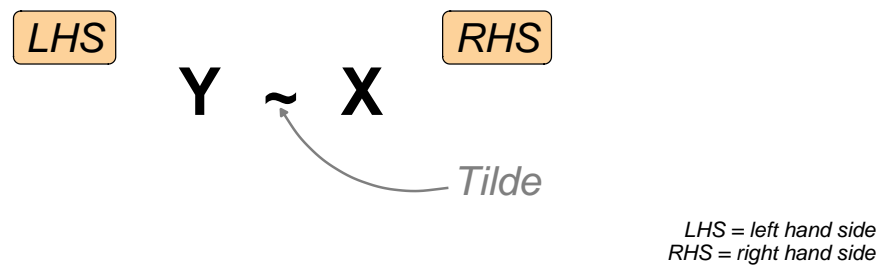


Figure 3.1: foo

Data, model, and error

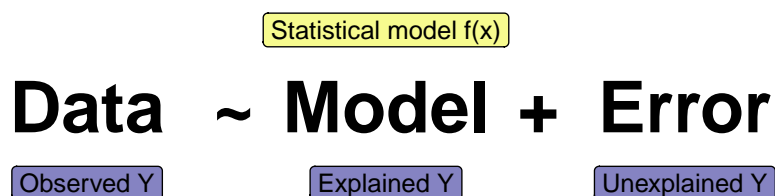


Figure 3.2: foo

```
[1] "#0D0887FF" "#6A00A8FF" "#B12A90FF" "#E16462FF" "#FCA636FF" "#F0F921FF"
```

```
[1] "#0D088780" "#6A00A880" "#B12A9080" "#E1646280" "#FCA63680" "#F0F92180"
```

Data, model, and error

Linear regression

$$\text{Data} \sim \text{Model} + \text{Error}$$

Observed Y Coefficients: β_0, β_1 Residuals

Figure 3.3: foo

Simple linear model

Outcome Intercept Slope Influencer

$$Y \sim \beta_0 + \beta_1 \cdot X_1 + \varepsilon$$

Coefficients Residuals

*Y = Outcome, response, endpoint or dependent variable
X = Influencer, explanator, predictor or independent variable*

Figure 3.4: foo

Multiple linear model

Outcome Intercept Slope Slope Influencer

$$Y \sim \beta_0 + \beta_1 \cdot X_1 + \dots + \beta_p \cdot X_p + \varepsilon$$

Coefficients Residuen

*Y = Outcome, response, endpoint or dependent variable
X = Influencer, explanator, predictor or independent variable*

Figure 3.5: foo

Model notation in R

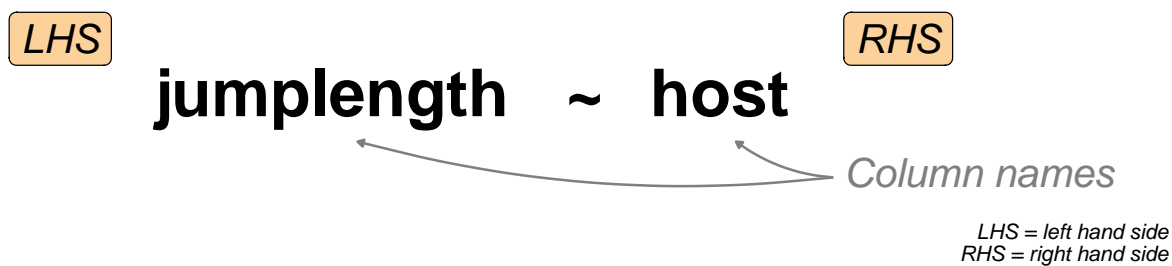


Figure 3.6: foo

Model notation in R

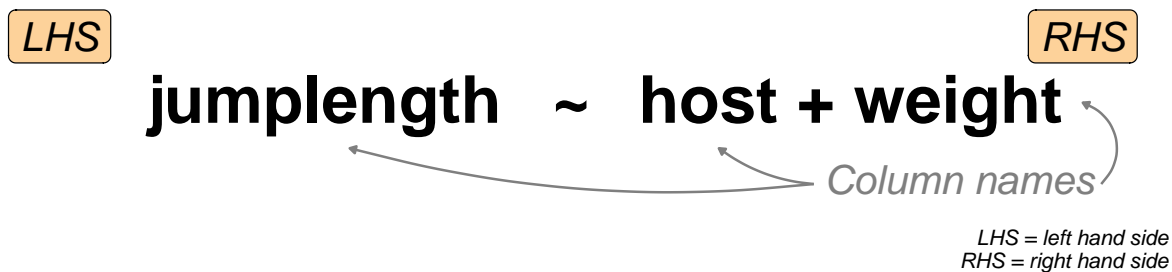


Figure 3.7: foo

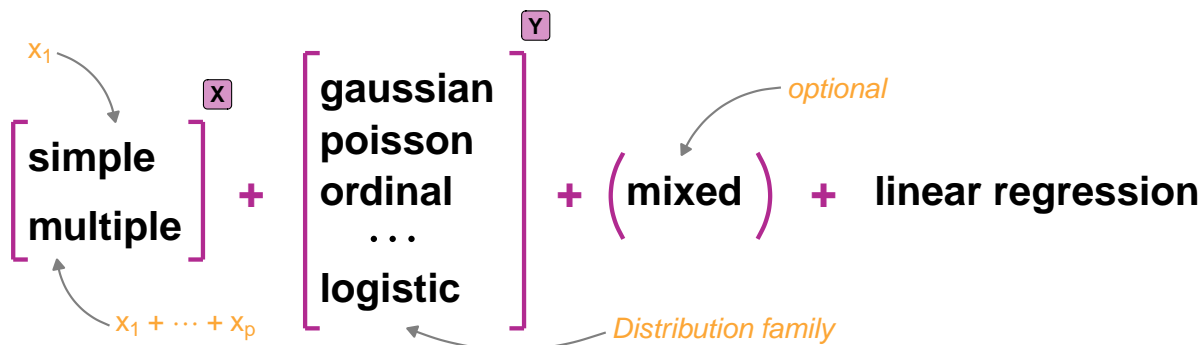


Figure 3.8: foo

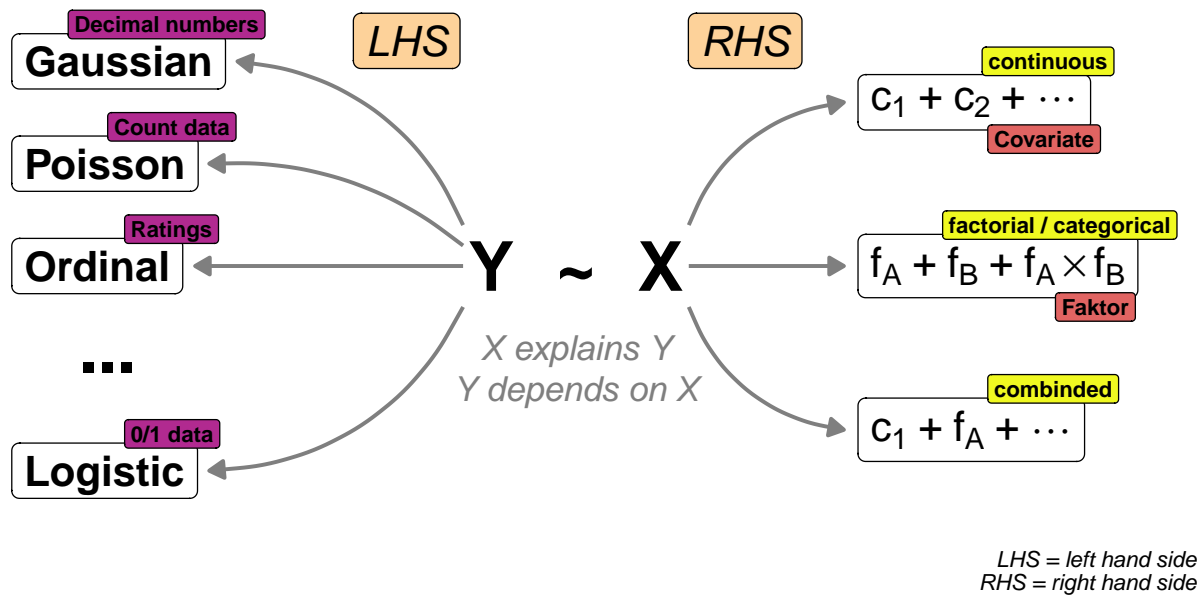


Figure 3.9: foo

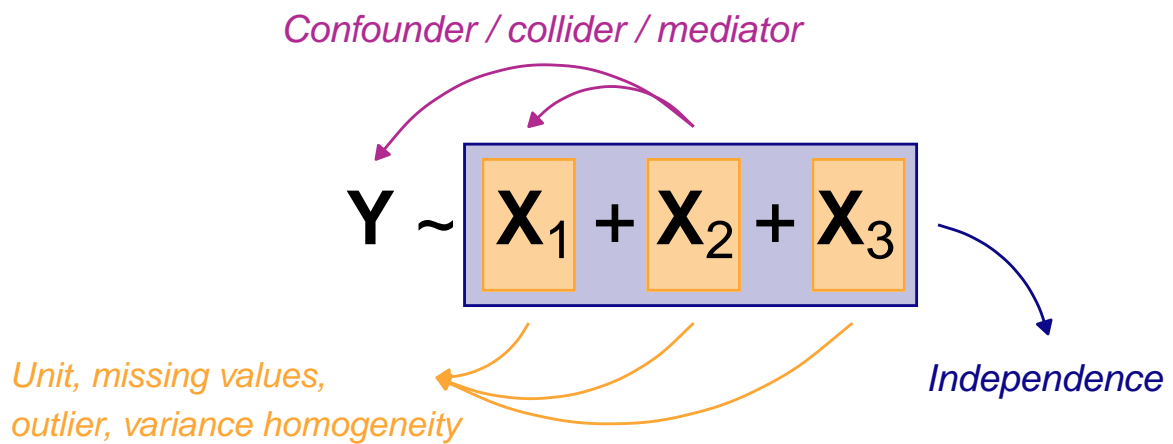


Figure 3.10: foo

3.3 R packages used

3.4 Data

3.5 Alternatives

Further tutorials and R packages on XXX

3.6 Glossary

term what does it mean.

3.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

3.8 Summary

References

Part II

Tales of data

Last modified on 22. November 2025 at 19:18:43

“X-Factor, the unfathomable. We live in a world where dreams and reality are closely intertwined, where facts often seem like figments of the imagination that we cannot explain. Can you distinguish between truth and lies? Yours, Jonathan Frakes.” — German intro of Beyond Belief: Fact or Fiction

Once upon a time Cadiergues and his team won the Nobel Prize in Stockholm for their work with dog and cat fleas⁴. This sparked a time of increased scientific research on fleas of all types of hosts, with scientists studying them in great detail. A wealth of research questions remained to be addressed, while funding opportunities lay in wait. We now have the great opportunity to use open-source data to revisit all the exciting findings in flea research. What are our main research topics and questions that we want to cover?

Unfortunately, none of this ever happened in our branch of reality. But it might happen in countless other branches⁵.

References

4 Body weight of fleas

[conflicted] Will prefer `dplyr::filter` over any other package.

Last modified on 23. November 2025 at 08:41:10

“A quote.” — Dan Meyer

4.1 General background

4.2 Theoretical background

4.3 R packages used

4.4 Data

4.4.1 Linear

```
jump_weight_tbl <- tibble(x = abs(rnorm(21, 3, 4)),  
                          y = 10 + 1.2 * x + rnorm(21, 0, 2)) |>  
  mutate_all(round, 1) |>  
  rename(weight = x, jumplength = y)
```

4.4.2 Non-linear

```
jump_weight_non_linear_tbl <- tibble(x = abs(rnorm(32, 3, 4)),  
                                     y = 0.15*x^3 - 2.2*x^2 + 8.8*x + 3.2 + rnorm(32, 0, 1))  
  rename(weight = x, jumplength = y)
```

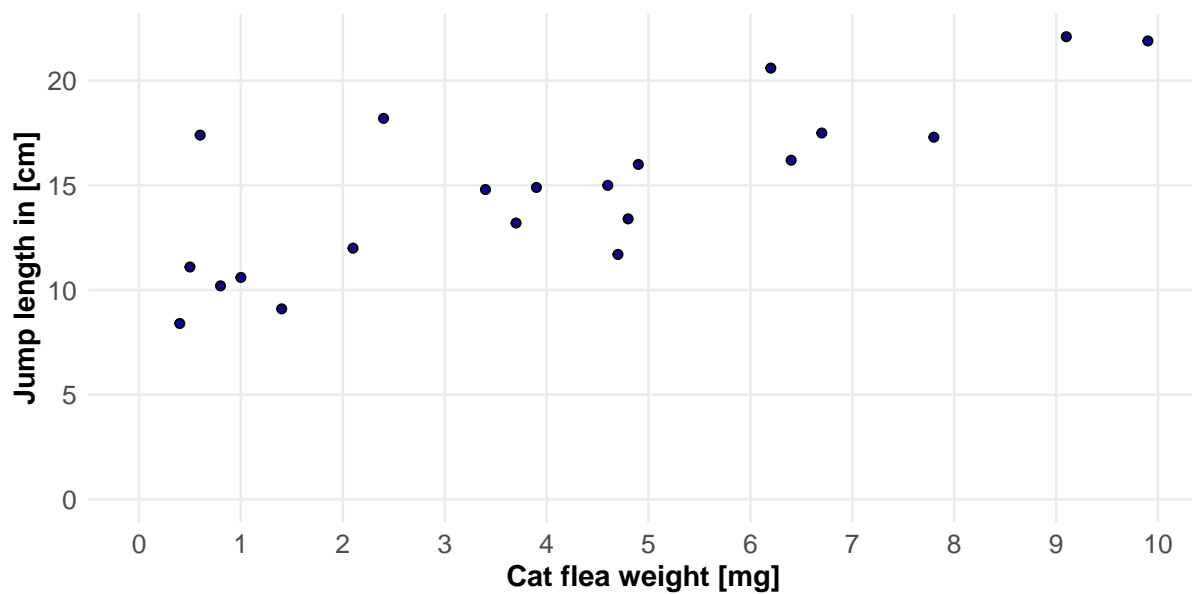


Figure 4.1: foo (A) foo (B) foo

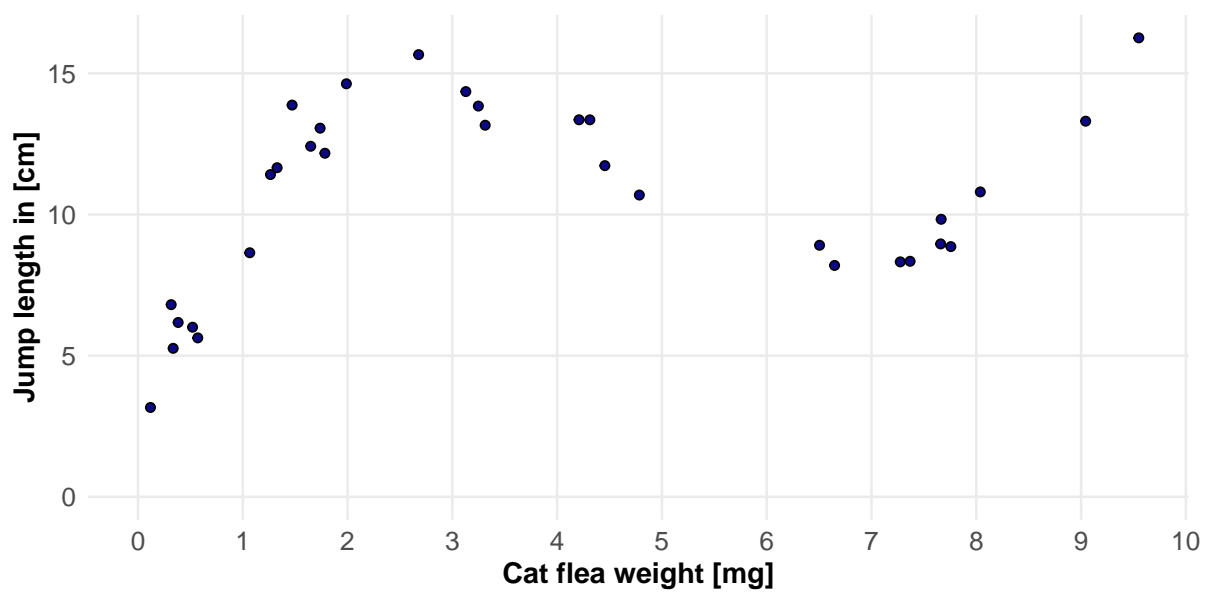


Figure 4.2: foo (A) foo (B) foo

4.5 Alternatives

Further tutorials and R packages on XXX

4.6 Glossary

term what does it mean.

4.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

4.8 Summary

References

5 Fleas of animals

Last modified on 07. December 2025 at 20:42:45

“A quote.” — Dan Meyer

5.1 General background

What is a data grid?

5.2 Theoretical background

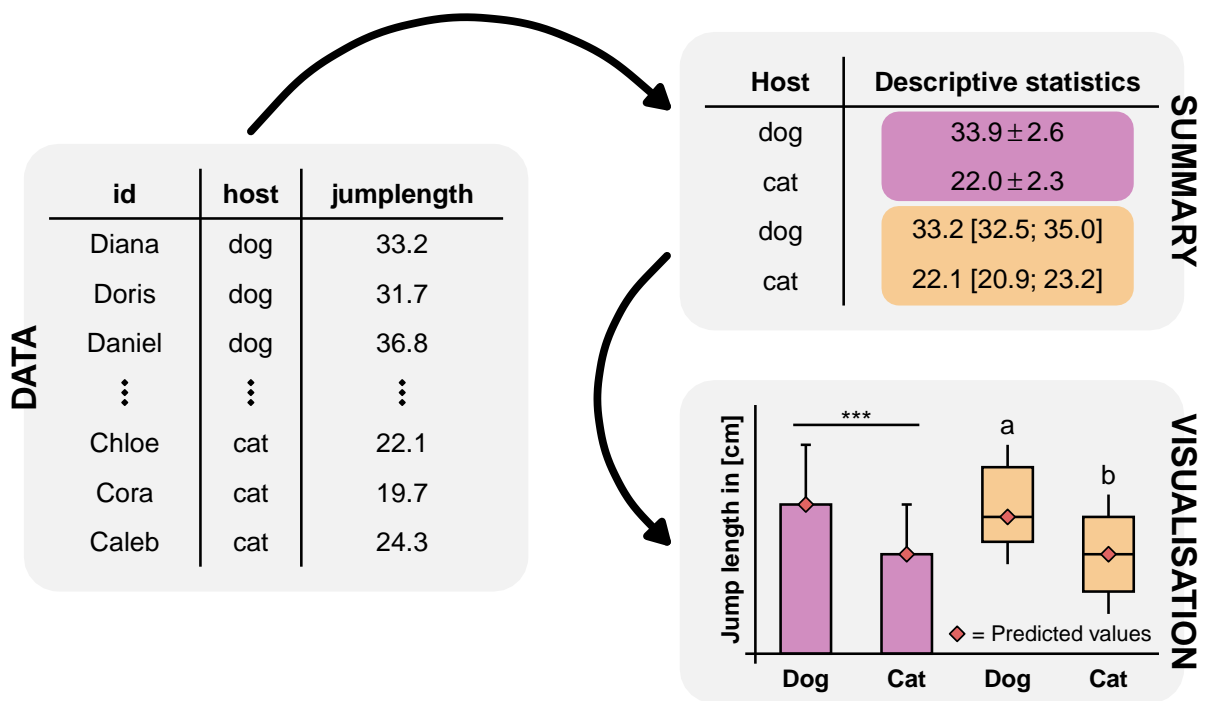


Figure 5.1: foo

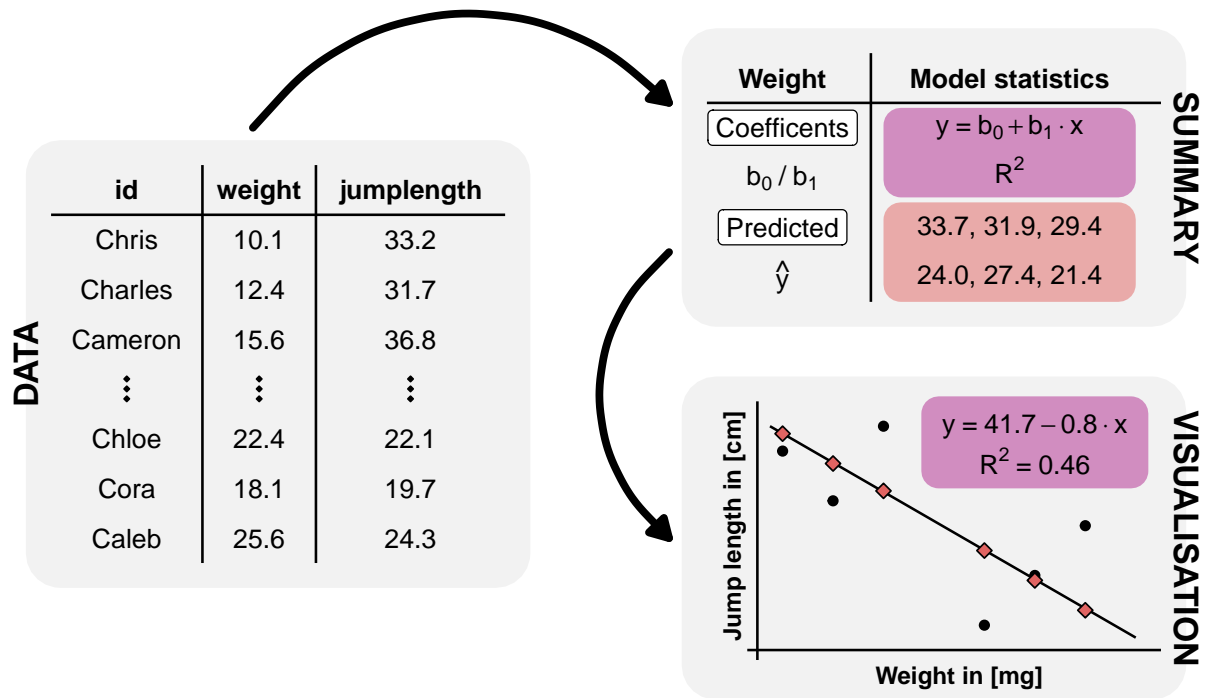


Figure 5.2: foo

5.3 R packages used

```
pacman::p_load(tidyverse, conflicted)
```

18

5.4 Data

Small data and grid

5.4.1 Jump performances of fleas

4

C. felis felis jump was $19.9 \pm 9.1\text{cm}$ with a range from 2 to 48cm

C. canis jump was longer $30.4 \pm 9.1\text{cm}$ with a range from 3 to 50cm

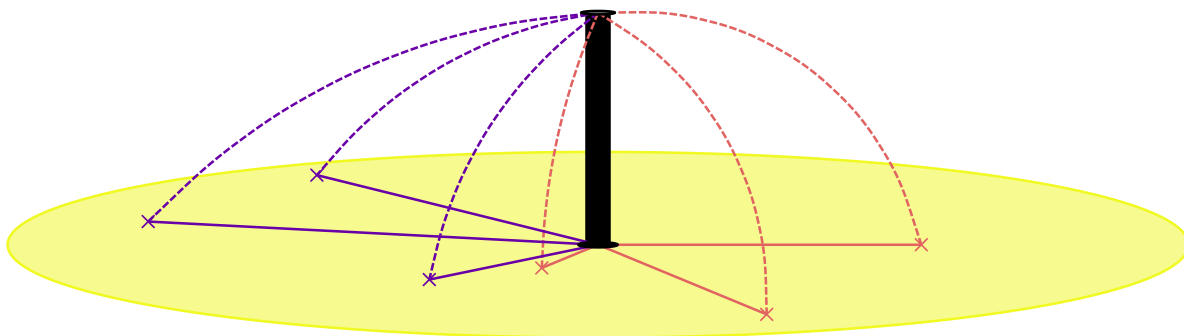


Figure 5.3: foo (A) foo (B) foo

```
jump_flea_grid <- expand_grid(host = c("cat", "dog")) |>
  mutate(mean = c(19.9, 30.4),
         sd = c(9.1))
```

```
jump_flea_tbl <- jump_flea_grid |>
  rowwise() |>
  mutate(jumplength = lst(rnorm(7, mean, sd))) |>
  unnest(cols = jumplength) |>
  mutate(host = as_factor(host))
```

```
jump_flea_tbl |>
  group_by(host) |>
  summarise(mean(jumplength),
            sd(jumplength)) |>
  mutate_if(is.numeric, round, 2)
```

```
# A tibble: 2 x 3
  host `mean(jumplength)` `sd(jumplength)`
  <fct>          <dbl>          <dbl>
1 cat             21.2             7.3
2 dog             33.0             6.57
```

```
jump_animals_grid <- expand_grid(host = c("cat", "fox", "rat", "dog")) |>
  mutate(mean = c(19.9, 35.2, 15.2, 30.4),
         sd = c(9.1, 10.3, 4.6, 9.1))
```

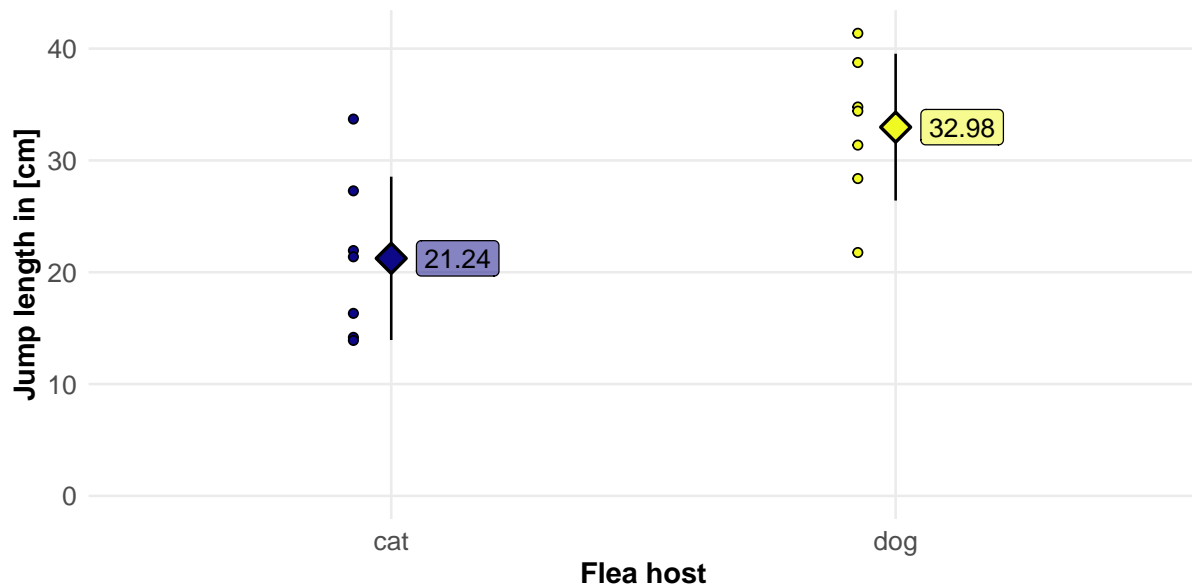



Figure 5.4: foo (A) foo (B) foo

```
jump_animals_tbl <- jump_animals_grid |>
  rowwise() |>
  mutate(jumplength = lst(rnorm(7, mean, sd))) |>
  unnest(cols = jumplength) |>
  mutate(host = as_factor(host))
```

```
jump_animals_tbl |>
  group_by(host) |>
  summarise(mean(jumplength),
            sd(jumplength)) |>
  mutate_if(is.numeric, round, 2)
```

```
# A tibble: 4 x 3
  host `mean(jumplength)` `sd(jumplength)`
  <fct>          <dbl>          <dbl>
1 cat             21.2             9.36
2 fox             34.6            10.2
3 rat             13.4             2.51
4 dog             34.0            17.4
```

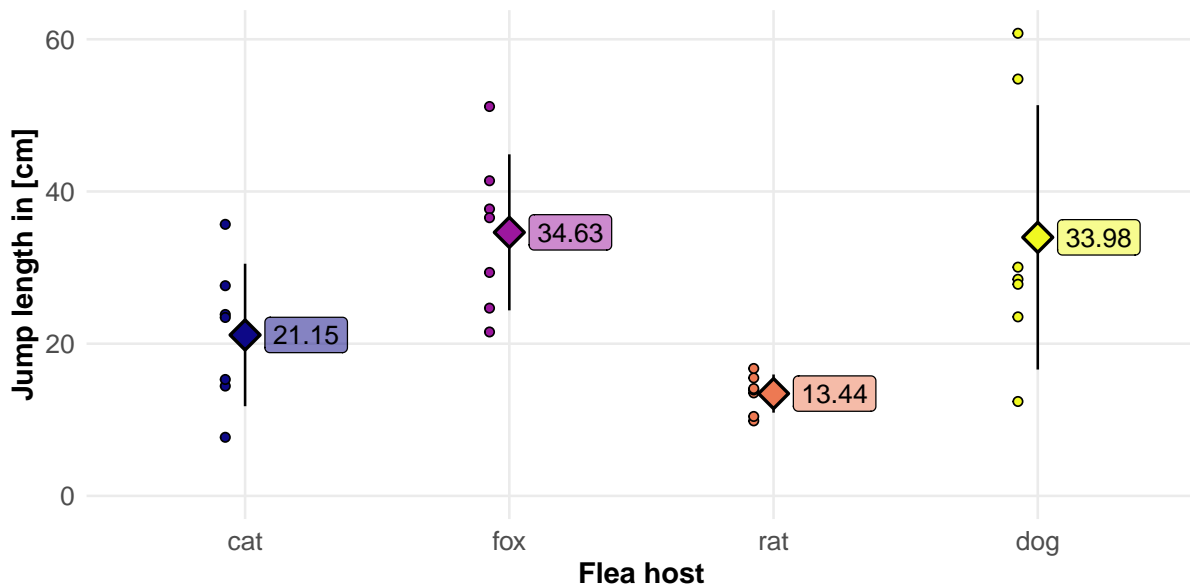


Figure 5.5: foo (A) foo (B) foo

5.4.2 Measurements on animal fleas

- Jump length in [cm] called `jump_length`
- Number of hairs on each leg called `count_leg_left` and `count_leg_right`
- Ratings of each flea, as listed in the catalog of the Fédération Internationale de la Beauté des Puces (FIBP) called `rating` on a Likert scale from 1 to 5, with 5 being the strongest expression.
- The infection status with the flea cold is called `infected`, with a value of 0/1 for no/yes.

```
tibble(jumplength = rnorm(7, 5, 1),
       countleg_left = rpois(7, 4),
       countleg_right = rpois(7, 4),
       countleg = (countleg_left + countleg_right)/2,
       rating = sample(1:5, 7, replace = TRUE, prob = c(0.1, 0.2, 0.4, 0.2, 0.1)),
       infectd = rbinom(7, prob = 0.5, size = 1))
```

A tibble: 7 x 6

	jumplength	countleg_left	countleg_right	countleg	rating	infectd
	<dbl>	<int>	<int>	<dbl>	<int>	<int>
1	6.17	4	8	6	3	1
2	5.03	5	5	5	4	0
3	6.51	6	1	3.5	3	0
4	4.38	2	1	1.5	3	1

5	2.62	0	3	1.5	3	0
6	6.64	3	3	3	2	0
7	3.82	3	3	3	3	0

5.4.3 Fleas in urban and rural habitats

Why so complex?

19

```
jump_habitat_grid <- expand_grid(host = 1:3, site = 1:2) |>
  mutate(mean_host = c(19.9, 19.9, 30.4, 30.4, 15.2, 15.2),
         mean_site = c(5, 0, 5, 0, 5, -5),
         mean = mean_host + mean_site,
         sd = c(9.1, 9.1, 9.1, 9.1, 4.6, 4.6))
jump_habitat_grid
```

```
# A tibble: 6 x 6
  host site mean_host mean_site mean sd
  <int> <int>   <dbl>   <dbl> <dbl> <dbl>
1     1     1     19.9       5  24.9  9.1
2     1     2     19.9       0  19.9  9.1
3     2     1     30.4       5  35.4  9.1
4     2     2     30.4       0  30.4  9.1
5     3     1     15.2       5  20.2  4.6
6     3     2     15.2      -5  10.2  4.6
```

```
jump_habitat_raw_tbl <- jump_habitat_grid |>
  rowwise() |>
  mutate(jumplength = lst(rnorm(7, mean, sd))) |>
  unnest(cols = jumplength)
```

```
jump_habitat_tbl <- jump_habitat_raw_tbl |>
  select(host, site, jumplength) |>
  mutate(host = factor(host, labels = c("cat", "dog", "rat")),
         site = factor(site, labels = c("urban", "rural"))) |>
  mutate_if(is.numeric, round, 2)
```

```
jump_habitat_tbl |>
  group_by(host, site) |>
  summarise(mean(jumplength),
            sd(jumplength)) |>
  mutate_if(is.numeric, round, 2)
```

A tibble: 6 x 4

Groups: host [3]

	host	site	`mean(jumplength)`	`sd(jumplength)`
	<fct>	<fct>	<dbl>	<dbl>
1	cat	urban	21.8	9.26
2	cat	rural	23.6	7.23
3	dog	urban	35.2	9.1
4	dog	rural	27.9	12.8
5	rat	urban	19.2	1.85
6	rat	rural	12.8	3.74

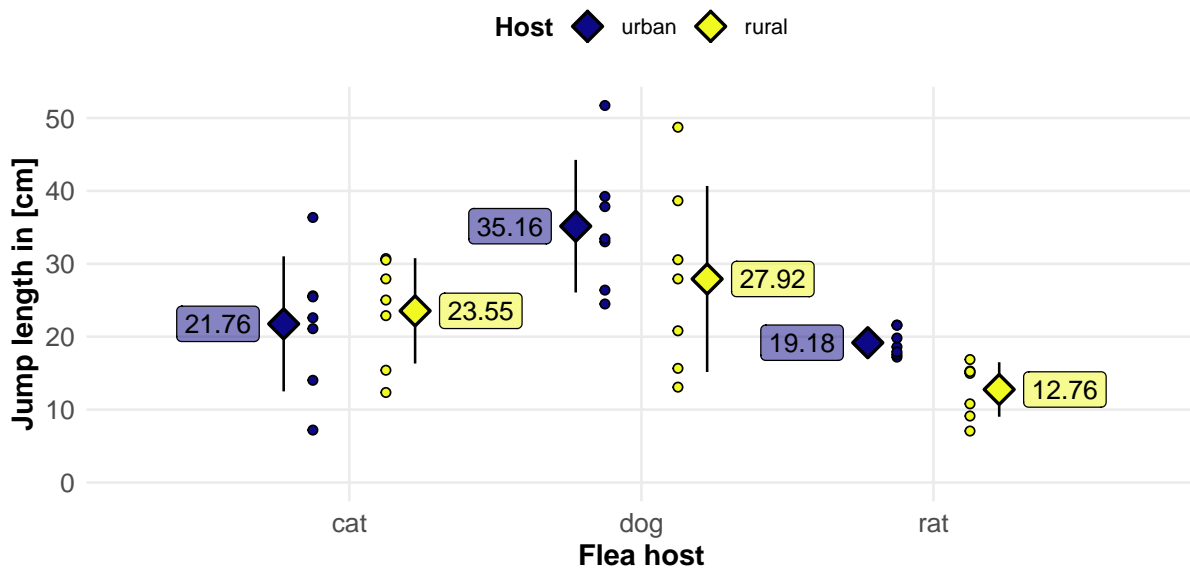


Figure 5.6: foo (A) foo (B) foo

5.5 Data availability

The data is available as txt-Files under <https://github.com/jkruppa/biodatascience>.

5.6 Alternatives

Further tutorials and R packages on XXX

5.6.1 Why linear is a bad idea?

```
jump_flea_tbl <- tibble(host = rep(0:1, each = 7),  
                        jump_length = 19.9 + 10.5 * host + rnorm(14, 0, 9.1)) |>  
  mutate(host = factor(host, labels = c("cat", "dog")))
```

5.6.2 Why is a tibble() bad idea?

```
jump_animals_tbl <- tibble(cat = rnorm(n = 7, mean = 19.9, sd = 9.1),  
                           fox = rnorm(n = 7, mean = 35.2, sd = 10.3),  
                           rat = rnorm(n = 7, mean = 15.2, sd = 4.6),  
                           dog = rnorm(n = 7, mean = 30.4, sd = 9.1)) |>  
  pivot_longer(cols = cat:dog, values_to = "jump_length", names_to = "host") |>  
  mutate(host = as_factor(host))
```

5.7 Glossary

term what does it mean.

5.8 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

5.9 Summary

References

6 Eating habits of fleas

[conflicted] Will prefer `dplyr::filter` over any other package.

Last modified on 25. November 2025 at 20:10:00

“A quote.” — Dan Meyer

6.1 General background

Based on standard methods in flea research and experimental entomology, as well as a similar published experiment, the following three types of food for adult cat fleas are possible, representing different nutritional conditions:

Blood: This is the natural and optimal food source for adult fleas. Often, defibrinated or anticoagulated animal blood (e.g. bovine blood, rabbit blood) is used for this purpose, which is offered in special in vitro feeding systems (e.g. through a membrane). Expectation: Fleas that are optimally nourished should show the greatest jumping distance.

Sugar water: Often serves as a ‘control feed’ or as a feed that provides energy (sugar) but lacks essential nutrients (such as proteins from blood). Expectation: Jumping distance could be reduced due to the lack of blood and thus the proteins important for reproduction, which could impair physiological fitness.

Ketchup - a nutrient-poor or unsuitable food: This option is used to simulate poor, incomplete or stressful nutritional conditions. In a similar documented experiment, ketchup (a combination of sugar, vinegar and minimal other substances, but no blood) was used as the third feed. Expectation: It can be assumed that fleas will show the shortest jumping distance under these conditions, as they lack both essential nutrients and the necessary energy.

6.2 Theoretical background

6.3 R packages used

6.4 Data

6.4.1 Linear

```
jump_weight_feeding_tbl <- expand_grid(f = c(0, 1),  
                                       x = abs(rnorm(21, 5, 4))) |>  
  mutate(y = 10 + 1.2 * x + 10 * f + rnorm(42, 0, 2),  
         f = factor(f, labels = c("sugar_water", "blood"))) |>  
  mutate_if(is.numeric, round, 1) |>  
  rename(weight = x, jumplength = y, feeding = f)
```

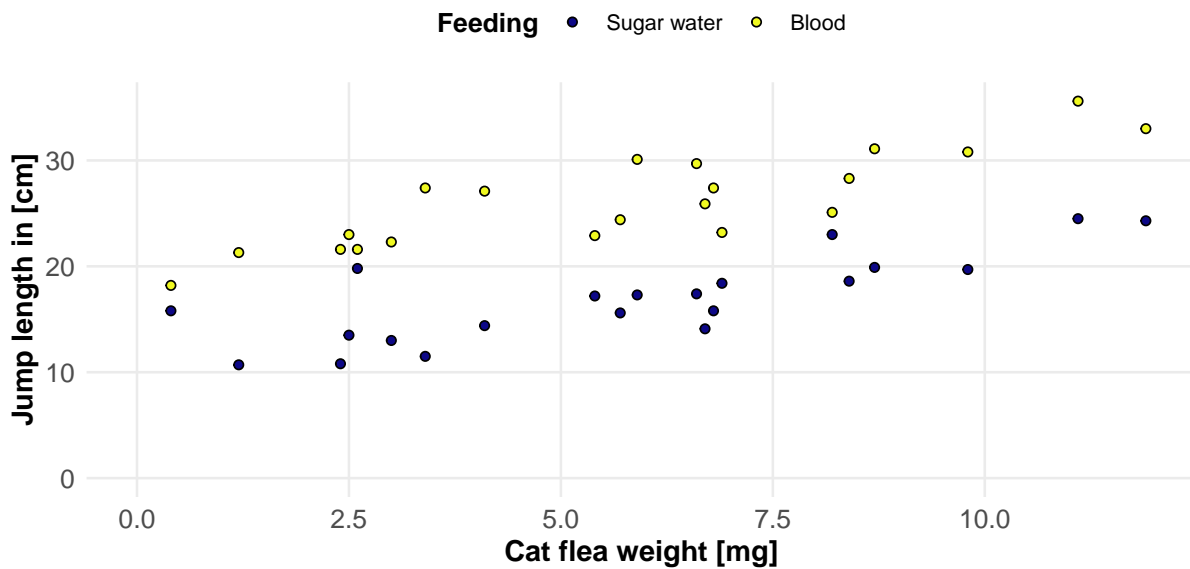


Figure 6.1: foo (A) foo (B) foo

6.4.2 Non linear

```
jump_weight_non_linear_tbl <- expand_grid(f = c(0, 1),
                                          x = abs(rnorm(42, 2.5, 2.5))) |>
  mutate(y = 25 - 4*f - 21 * exp(-0.2 * x^2 - 1.1 * f) + rnorm(84, 0, 1),
         f = factor(f, labels = c("sugar_water", "blood"))) |>
  mutate_if(is.numeric, round, 1) |>
  rename(weight = x, jumplength = y, feeding = f)
```

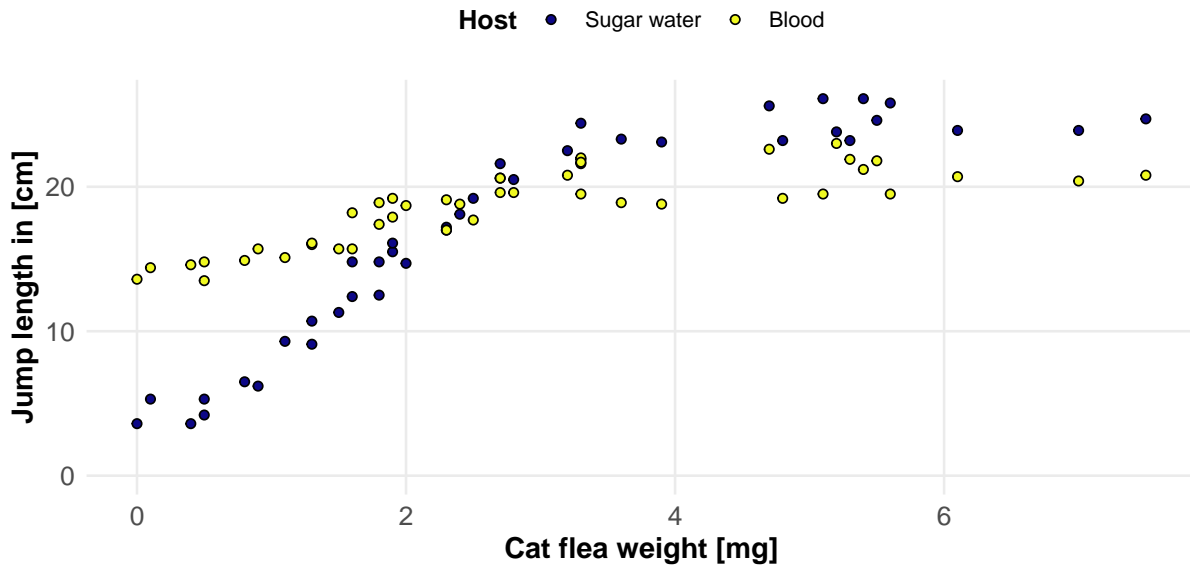


Figure 6.2: foo (A) foo (B) foo

6.4.3 International study

Degree of Urbanization of the United nations (rural as low-dense areas, town as semi-dense areas, cities as high-dense areas)

ODer das ganze mal mit Bodylength?

```
n_obs <- 41
jump_international_grid <- expand_grid(country = 1:9) |>
  mutate(mean_country = rnorm(9, 0, 4)) |>
  expand_grid(site = 1:3) |>
  mutate(mean_site = rnorm(27, 0, 4)) |>
  expand_grid(rep = 1:n_obs) |>
  mutate(bodyweight = abs(rnorm(n = (9*3*n_obs), 2.5, 2.5)),
         jumplength = 10 + 1.5 * bodyweight + mean_country + mean_site + rnorm(n = (9*3*n_obs), 0, 1))
```

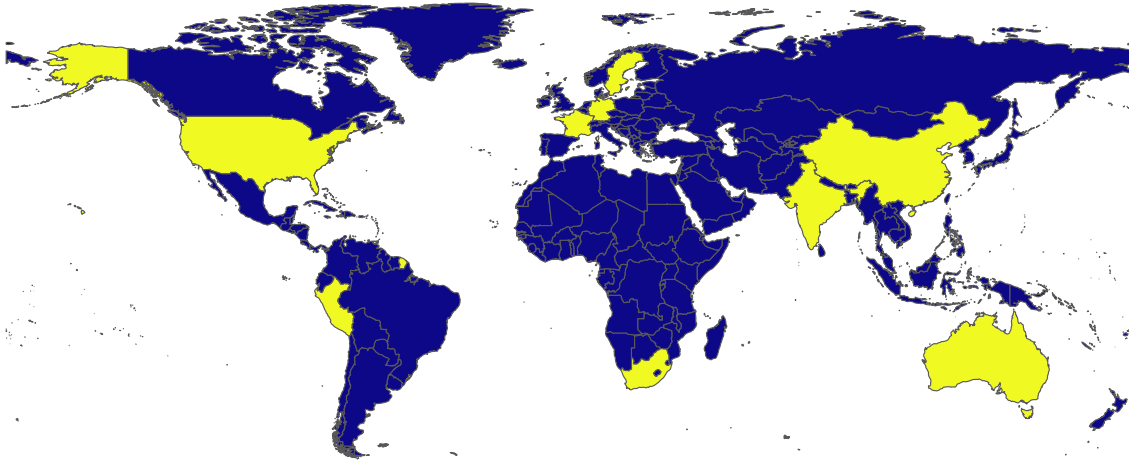



Figure 6.3: foo (A) foo (B) foo

DE = Germany FR = France IN = India US = United States CN = China AU = Australia
 PE = Peru SE = Sweden ZA = South Africa

```
jump_international_tbl <- jump_international_grid |>
  select(bodyweight, country, site, jumplength) |>
  mutate(country = factor(country, labels = c("DE", "FR", "AU",
                                             "IN", "US", "CN",
                                             "PE", "SE", "ZA")),
         site = factor(site, labels = c("rural", "semi-dense", "city"))) |>
  mutate_if(is.numeric, round, 2)
```

```
library(lme4)
```

Warning: package 'lme4' was built under R version 4.4.1

Loading required package: Matrix

Warning: package 'Matrix' was built under R version 4.4.1

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

expand, pack, unpack

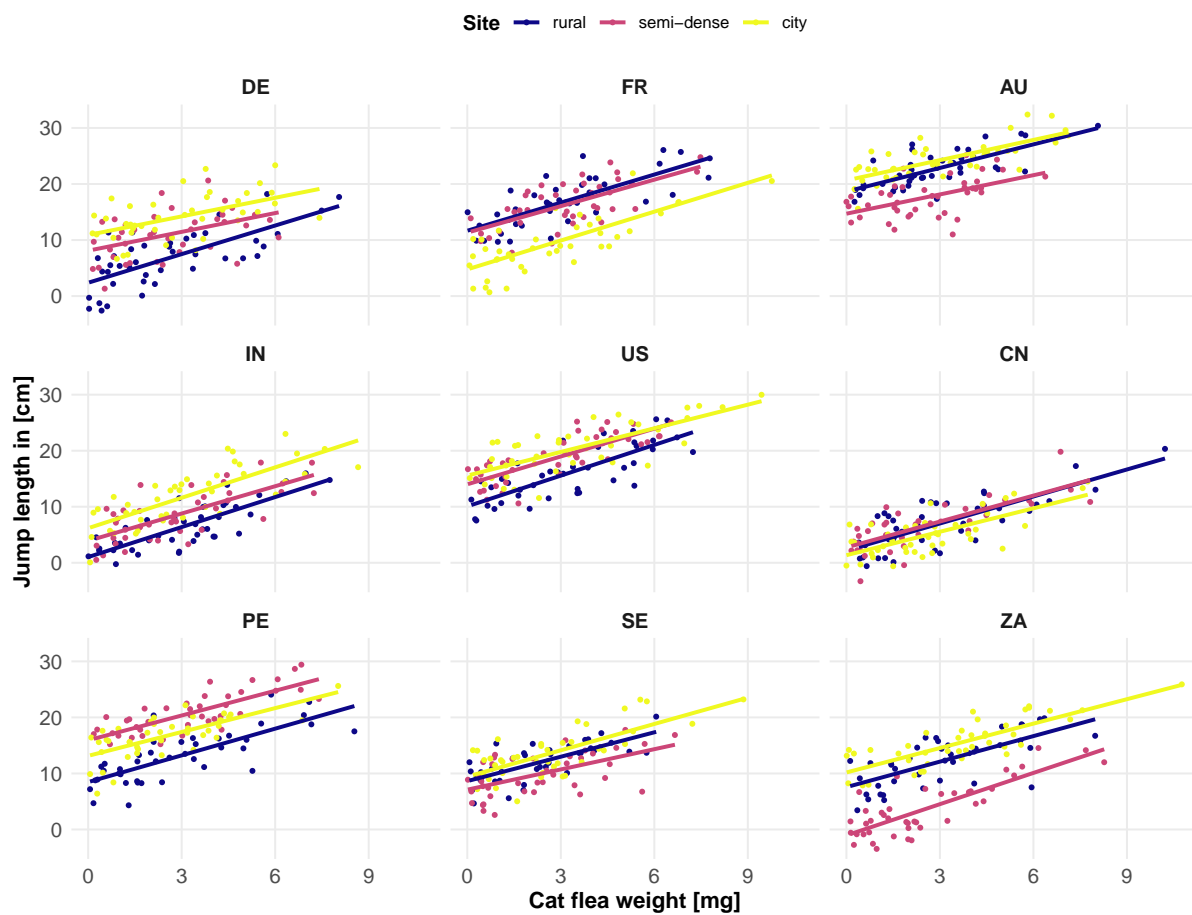


Figure 6.4: foo (A) foo (B) foo

```
lmer(jumplength ~ bodyweight + (site|country), data = jump_international_tbl)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: jumplength ~ bodyweight + (site | country)
Data: jump_international_tbl
REML criterion at convergence: 5640.55
Random effects:
Groups   Name                Std.Dev. Corr
country  (Intercept)            5.321
          sitesemi-dense    4.306   -0.34
          sitecity          4.251   -0.35  0.36
Residual                    2.926
Number of obs: 1107, groups:  country, 9
Fixed Effects:
(Intercept)  bodyweight
      8.811      1.532
```

```
c("#0D0887FF", "#5402A3FF", "#8B0AA5FF", "#B93289FF",
  "#DB5C68FF", "#F48849FF", "#FEBC2AFF", "#F0F921FF")
```

```
[1] "#0D0887FF" "#5402A3FF" "#8B0AA5FF" "#B93289FF" "#DB5C68FF" "#F48849FF"
[7] "#FEBC2AFF" "#F0F921FF"
```

6.5 Alternatives

Further tutorials and R packages on XXX

```
expand_grid(obs = c(1, 3, 2)) |>
  rowwise() |>
  mutate(foo = list(expand_grid(1:obs))) |>
  unnest(cols = c(foo))
```

```
# A tibble: 6 x 2
  obs `1:obs`
  <dbl>   <int>
1     1     1
2     3     1
3     3     2
4     3     3
5     2     1
6     2     2
```

6.6 Glossary

term what does it mean.

6.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

6.8 Summary

References

Part III

Speaking to data

Last modified on 25. October 2025 at 20:17:25

“What problem have you solved, ever, that was worth solving where you knew all the given information in advance? No problem worth solving is like that. In the real world, you have a surplus of information and you have to filter it, or you don’t have sufficient information and you have to go find some.” — [Dan Meyer in Math class needs a makeover](#)

Here comes the preface text

7 Programing in the 21st century

[conflicted] Will prefer `dplyr::filter` over any other package.

Last modified on 20. November 2025 at 14:02:31

“A quote.” — Dan Meyer

7.1 General background

20

18

21

22

7.2 Theoretical background

7.3 R packages used

7.4 Data

7.5 Alternatives

Further tutorials and R packages on XXX

7.6 Glossary

term what does it mean.

7.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

7.8 Summary

References

Part IV

Hypotheses testing

Last modified on 27. November 2025 at 15:17:37

“What problem have you solved, ever, that was worth solving where you knew all the given information in advance? No problem worth solving is like that. In the real world, you have a surplus of information and you have to filter it, or you don’t have sufficient information and you have to go find some.” — [Dan Meyer in Math class needs a makeover](#)

Here comes the preface text

8 Statistical testing

Last modified on 08. December 2025 at 19:07:26

“A quote.” — Dan Meyer

8.1 General background

“Statistik ist: Wenn der Jäger am Hasen einmal links und einmal rechts vorbeischießt, dann ist der Hase im Durchschnitt tot.” — Mike Krüger, German comedian

Statistics is: When the hunter misses the rabbit once to the left and once to the right, on average the rabbit is dead.

Statistics means: If the hunter misses the rabbit once to the left and once to the right, then on average, the rabbit’s a goner.

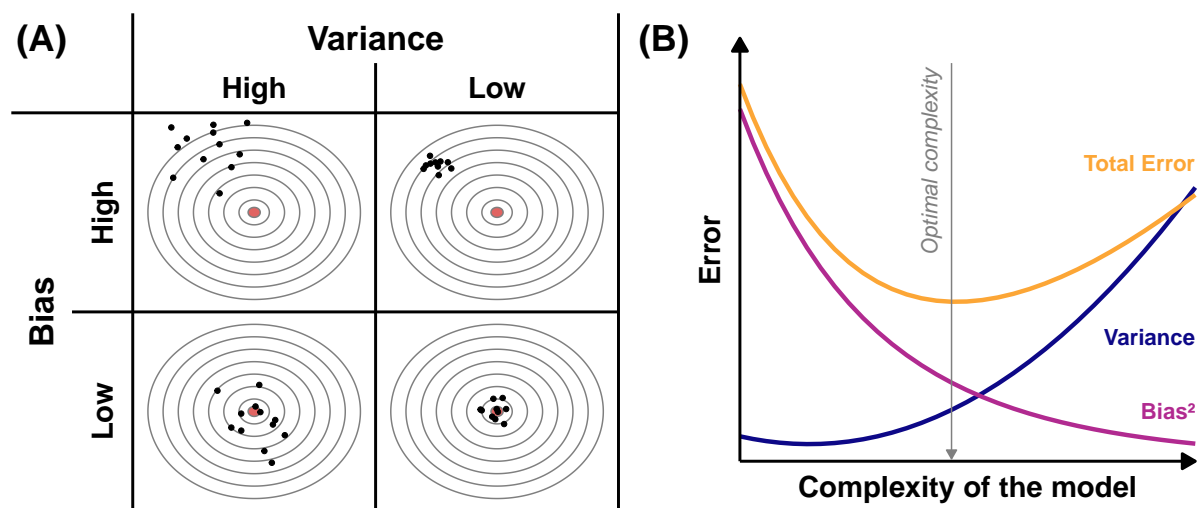


Figure 8.1: foo

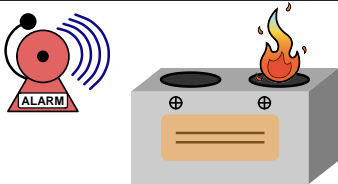
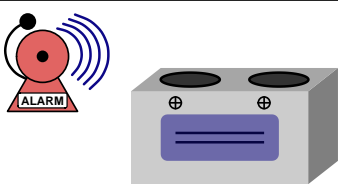
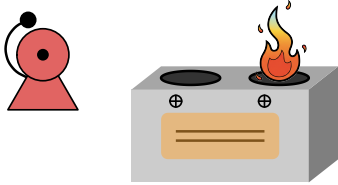
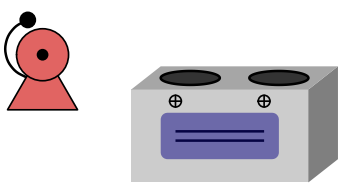
		Unkown truth in the population	
		Null is false / effect	Null is true / no effect
Decision by a statistical test	Reject Null / effect	 <p><i>ALARM with FIRE</i></p> <p>Power / True positive $1 - \beta = 80\%$ <i>We believe in this to be high.</i></p>	 <p><i>ALARM without FIRE</i></p> <p>Type I error / False positive $\alpha = 5\%$ <i>We control this in our testing.</i></p>
	Keep Null / no effect	 <p><i>FIRE without ALARM</i></p> <p>Type II error / False negative $\beta = 20\%$ <i>We cannot control this in our testing.</i></p>	 <p><i>No FIRE, no ALARM</i></p> <p>True negative $1 - \alpha = 95\%$ <i>We are rarely interested in.</i></p>

Figure 8.2: foo

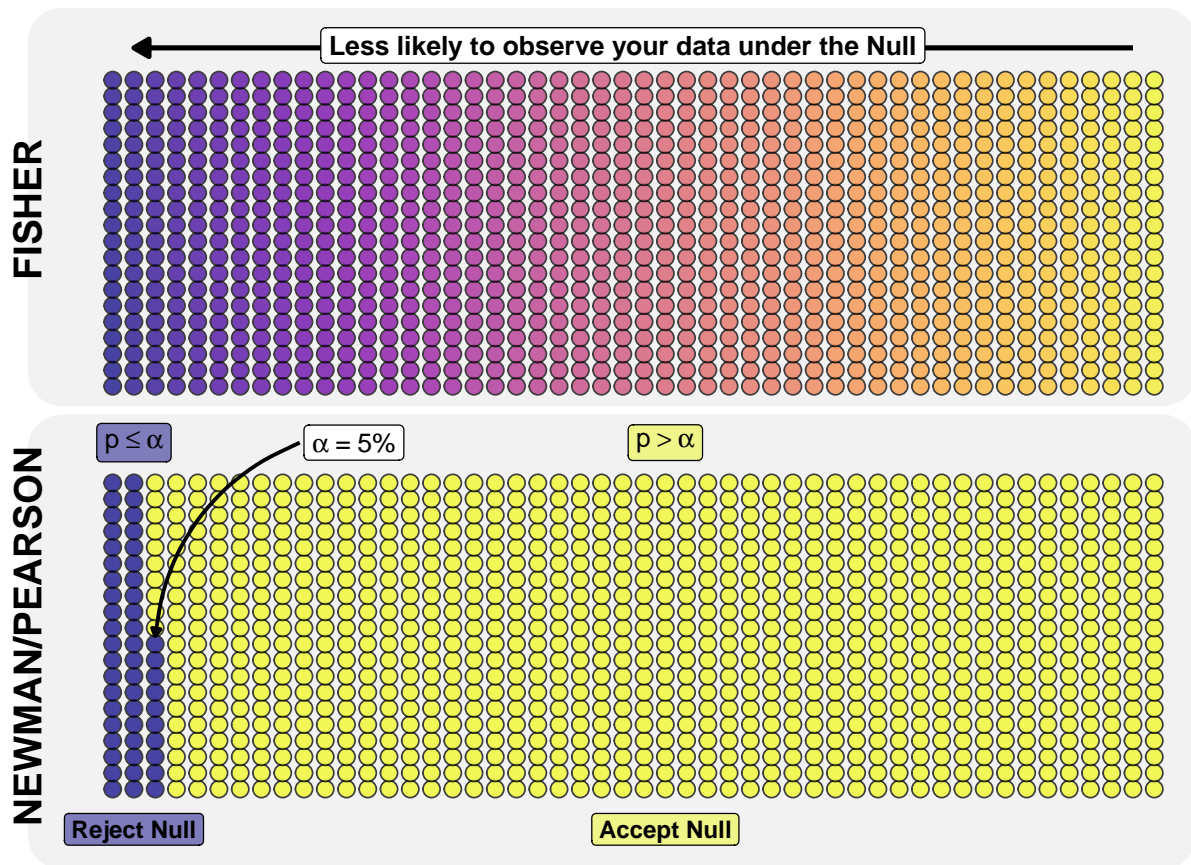


Figure 8.3: foo

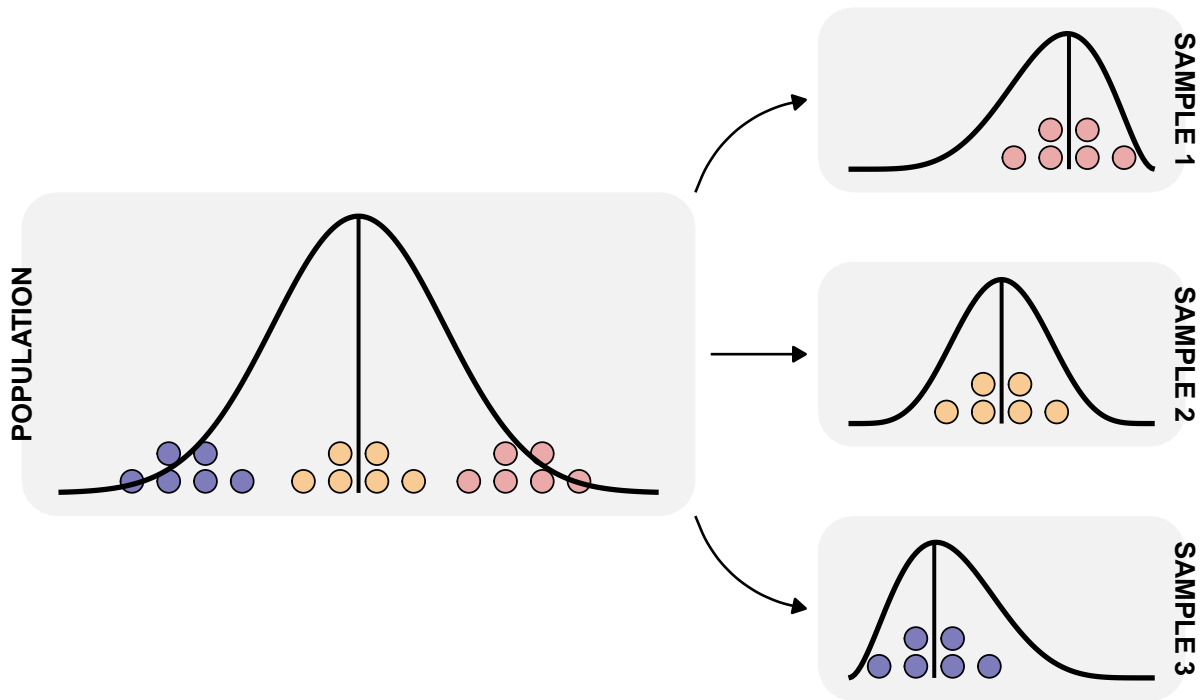


Figure 8.4: foo

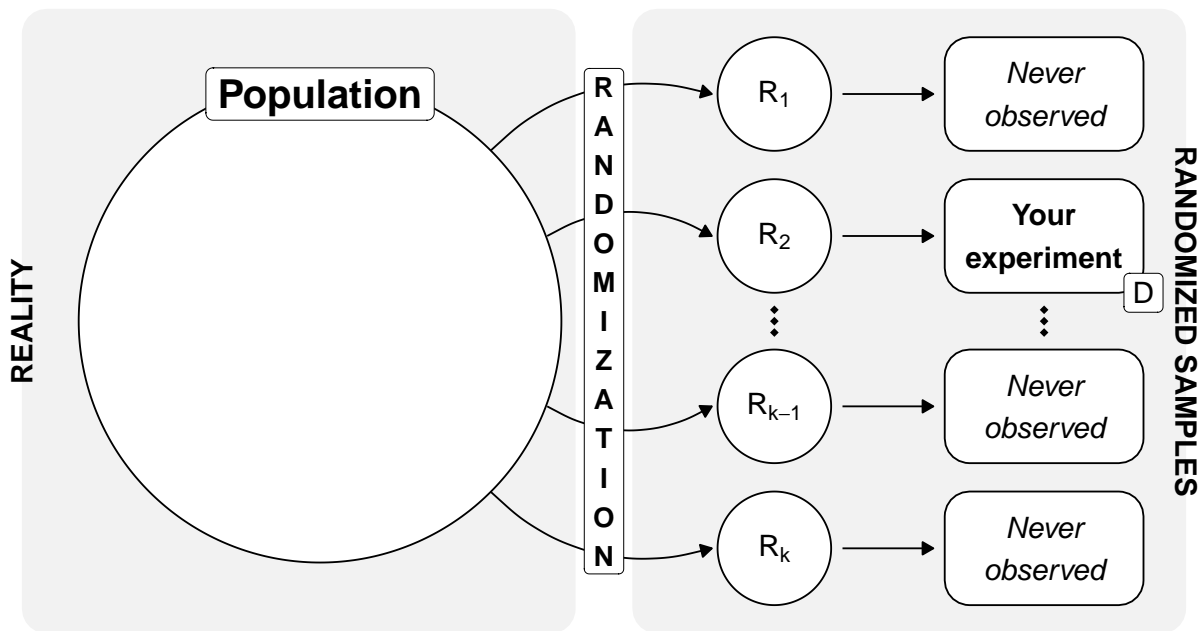


Figure 8.5: foo

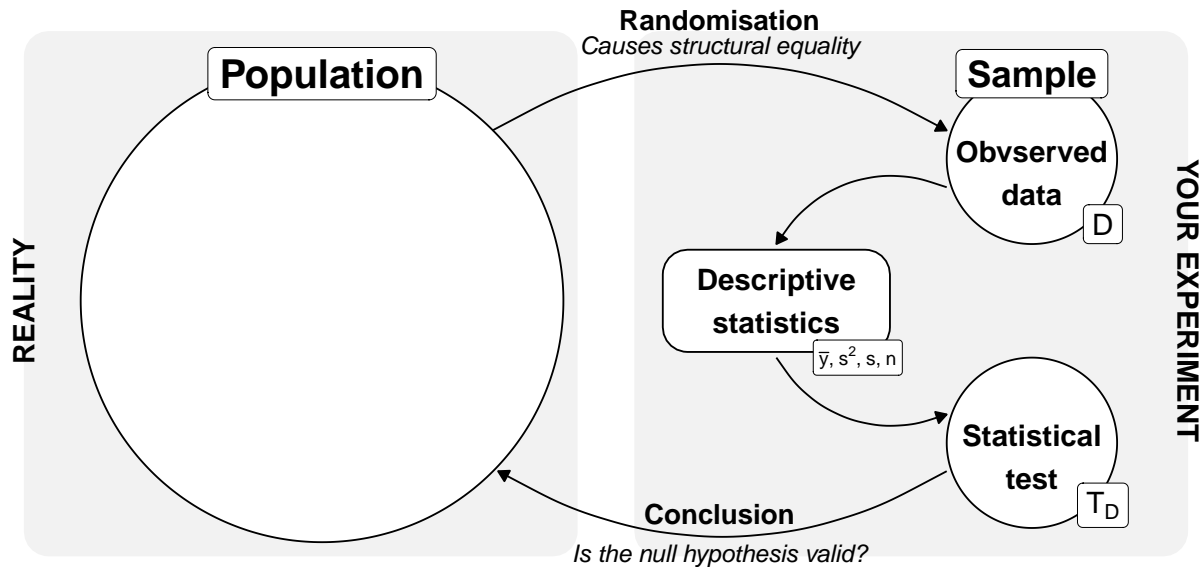


Figure 8.6: foo

8.2 Theoretical background

8.2.1 Fisher's approach: The 'significance test'

Fisher saw statistics as a tool for inductive reasoning (learning from data for science).

- Only ONE hypothesis: There is only the null hypothesis (H_0 e.g., 'no effect'). An alternative does not formally exist.
- The measure (p-value): The p-value is a continuous measure of the strength of evidence against the null hypothesis H_0 .
 - $p = 0.01$ Strong evidence against the null hypothesis.
 - $p = 0.20$ No evidence against the null hypothesis.
- The result: One rejects the null hypothesis H_0 or one does not make a judgement. One never 'accepts' the null hypothesis (one simply has not found enough evidence to reject it).
- Objective: Gain knowledge through individual experiments.

8.2.2 Neyman-Pearson's approach: The 'hypothesis test'

Neyman and Pearson sharply criticised Fisher. They said, 'You can't reject anything if you don't know what to accept instead.' They saw statistics as a decision-making process (behaviourism).

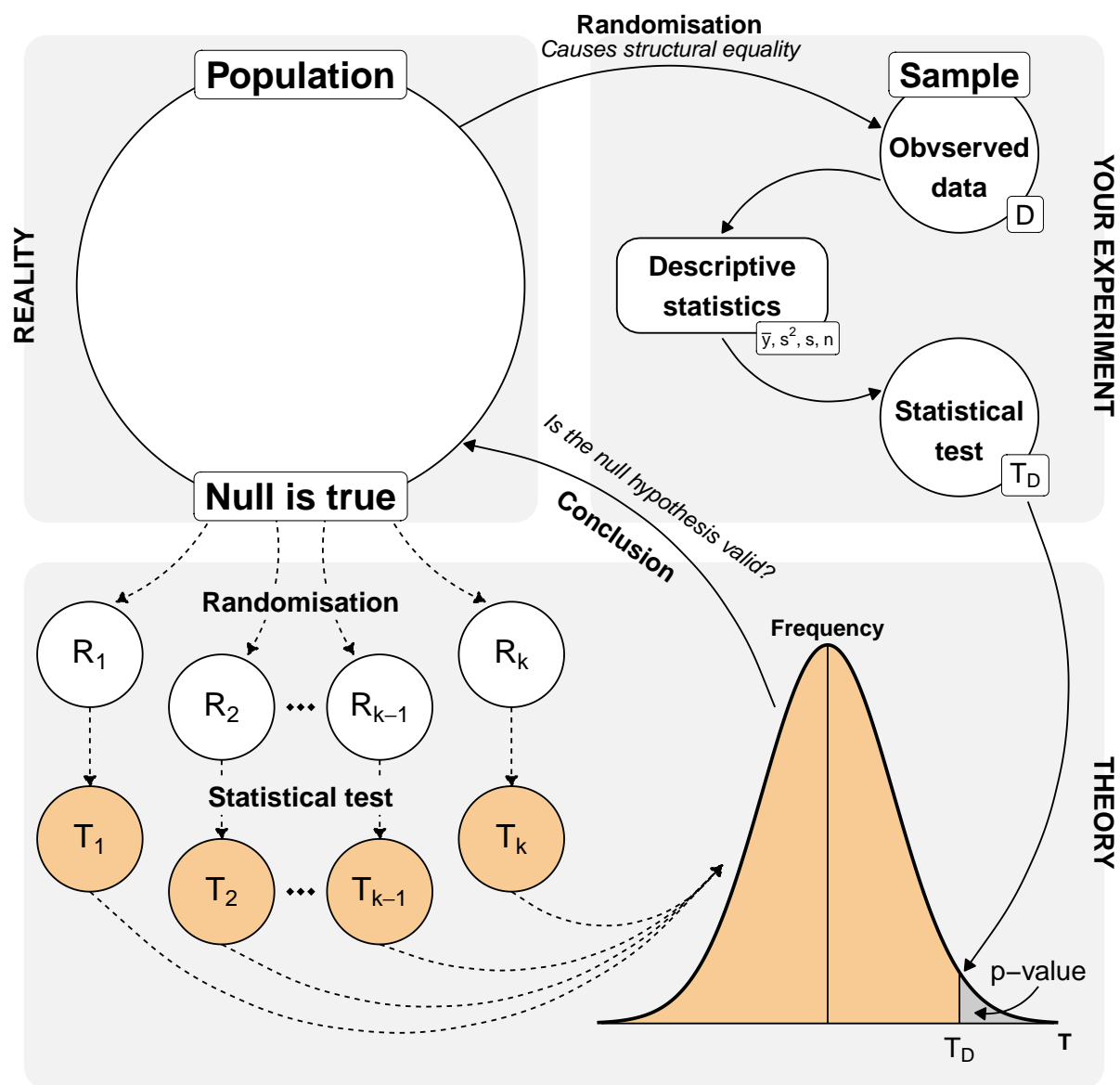


Figure 8.7: foo

- TWO hypotheses: There is the null hypothesis (H_0) AND a specific alternative hypothesis (H_A).
- Type 1 and 2 errors: Before the experiment begins, the following is determined:
 - α (alpha): How often am I allowed to incorrectly find an effect? (e.g. 5%)
 - β (Beta): How often am I allowed to mistakenly overlook a real effect? (Power/test strength).
- The result: A tough decision. ‘Accept H_0 ’ or ‘Reject H_0 ’ (or Accept H_A).
- Goal: Minimisation of losses over many repeated experiments (as in industrial production).

Neymans philosophy: We are not looking for the ‘truth’ in individual cases, but rather we behave in such a way that we are wrong as rarely as possible in 1000 decisions.

8.2.3 Today’s ‘hybrid chaos’

Modern textbooks and software (such as SPSS or R) often use a hybrid that historically makes no sense:

- We define $\alpha = 5\%$ (Neyman-Pearson).
- We calculate an exact p-value (Fisher).
- We report the p-value as evidence (Fisher), but use it for a hard yes/no decision (Neyman-Pearson).
- We talk about ‘power’ (Neyman-Pearson), but often only test against a non-specific alternative.

This mishmash often leads to misunderstandings, such as that a $p = 0.001$ indicates a ‘stronger effect’ than $p = 0.049$ (Fisher thinking), even though in Neyman-Pearson logic at $\alpha = 5\%$, one would have to make exactly the same decision in both cases (‘Reject H_0 ’).

8.3 R packages used

8.4 Data

8.5 Alternatives

Further tutorials and R packages on XXX

8.6 Glossary

term what does it mean.

8.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

8.8 Summary

References

Part V

Template Preface

Last modified on 25. October 2025 at 20:17:25

“What problem have you solved, ever, that was worth solving where you knew all the given information in advance? No problem worth solving is like that. In the real world, you have a surplus of information and you have to filter it, or you don’t have sufficient information and you have to go find some.” — [Dan Meyer in Math class needs a makeover](#)

Here comes the preface text

9 Template chapter

[conflicted] Will prefer `dplyr::filter` over any other package.

Last modified on 17. November 2025 at 20:02:35

“A quote.” — Dan Meyer

9.1 General background

9.2 Theoretical background

9.3 R packages used

9.4 Data

9.5 Alternatives

Further tutorials and R packages on XXX

9.6 Glossary

term what does it mean.

9.7 The meaning of “Models of Reality” in this chapter.

- itemize with max. 5-6 words

9.8 Summary

References

A Why does it look like this?

[conflicted] Will prefer `dplyr::filter` over any other package.

Last modified on 20. November 2025 at 10:01:22

“A quote.” — Dan Meyer

A.1 Used R packages

This is a printout of the `init.R` file, which can be found on GitHub

```
pacman::p_load(tidyverse, ggforce, viridis, knitr, patchwork)
```

A.2 Used theme of the visualisations

```
theme_book <- function() {  
  theme_minimal() +  
    theme(panel.grid.minor = element_blank(),  
          plot.title = element_text(size = 16, face = "bold"),  
          plot.subtitle = element_text(size = 12, face = "italic"),  
          plot.caption = element_text(face = "italic"),  
          axis.title = element_text(size = 12, face = "bold"),  
          axis.text = element_text(size = 11),  
          legend.title = element_text(face = "bold"))  
}
```

A.3 Used color palettes

```
col_pal <- \(\n, alpha) plasma(n = n, alpha = alpha)
c6_pal <- col_pal(6, 1)
```

References

- [1] Smolin L. *Einstein's Unfinished Revolution: The Search for What Lies Beyond the Quantum*. Penguin; 2019.
- [2] Feynman RP, Leighton R. " *Surely You're Joking, Mr. Feynman!*": *Adventures of a Curious Character*. Random House; 1992.
- [3] Feynman RP. Cargo cult science. In: *The Art and Science of Analog Circuit Design*. Elsevier; 1998:55-61.
- [4] Cadiergues MC, Joubert C, Franc M. A comparison of jump performances of the dog flea, *Ctenocephalides canis* (Curtis, 1826) and the cat flea, *Ctenocephalides felis felis* (Bouché, 1835). *Veterinary parasitology*. 2000;92(3):239-241.
- [5] Everett III H. 'Relative state' formulation of quantum mechanics. *Reviews of modern physics*. 1957;29(3):454.
- [6] Clarke AC. Clarke's third law on UFO's. *Science*. 1968;159(3812):255-255.
- [7] Chalmers AF. *What Is This Thing Called Science?* Hackett Publishing; 2013.
- [8] Feynman R. What is science. Published online 1966.
- [9] Campbell NR. What is science? Published online 1952.
- [10] Clark M, Berry S. *Models Demystified: A Practical Guide from Linear Regression to Deep Learning*. CRC Press; 2025.
- [11] Cox D, Efron B. Statistical thinking for 21st century scientists. *Science advances*. 2017;3(6):e1700768.
- [12] Deutsch D. *The Beginning of Infinity: Explanations That Transform the World*. Penguin UK; 2011.
- [13] McCullagh P. What is a statistical model? *The Annals of Statistics*. 2002;30(5):1225-1310.

- [14] Appleton DR. What do we mean by a statistical model? *Statistics in medicine*. 1995;14(2):185-197.
- [15] Hand D. What is the purpose of statistical modeling? *Harvard Data Science Review*. 2019;1(1).
- [16] Spanos A. Where do statistical models come from? Revisiting the problem of specification. *Lecture Notes-Monograph Series*. Published online 2006:98-119.
- [17] Gilchrist W. *Statistical Modelling*. Wiley Chichester; 1984.
- [18] Wickham H, Çetinkaya-Rundel M, Golemund G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. " O'Reilly Media, Inc."; 2023.
- [19] Wickham H, Henry L. *Purrr: Functional Programming Tools*.; 2025. <https://purrr.tidyverse.org/>
- [20] Dormann CF, Ellison AM. *Statistics by Simulation: A Synthetic Data Approach*. Princeton University Press; 2025.
- [21] Wickham H. *Conflicted: An Alternative Conflict Resolution Strategy*.; 2023. <https://conflicted.r-lib.org/>
- [22] Rinker T, Kurkiewicz D. *Pacman: Package Management Tool*.; 2012. <https://github.com/trinker/pacman>