Models of Reality

The Illustrative Guide

Prof. Dr. Jochen Kruppa-Scheetz

2025-12-09

Table of contents

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.

I need to see it. Therefore, this book is a visualization whenever it is possible or feasible. Some more complex models will be described rather than visualized.

There are better mathematical or theoretical books. I would recommend the book by Westfall and Arias (2020)[1]. It is an enjoyable read and provides a wealth of information on regression models. The visuals are mediocre at best. The R code is outdated and not state of the art for the 21st century.

Some ideas are not new. I will present them in a different way. I think this is an effective approach to writing a book. The book aims to teach you and enable you to teach statistics to beginners. It may also be beneficial for more advanced classes, but my emphasis was always on the uninformed learners.

Lying to children or some angels will be hurt. Sometimes math is a shield to hide behind. Not really the math itself but the mathematical notations. If you cant speak the math language you cannot follow the train of thought. Because communication the the search for compliance, misunderstanding a language can be barrier.

Machine room of science

Bloody beginners and the curse of knowledge

> *"To write a book you must become the book." — Naval Ravikant*

> *"Ludwig Boltzman, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics. Perhaps it will be wise to approach the subject cautiously." — Opening lines of 'States of Matter', by D.L. Goodstein.*

References

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.

text[2]

> *“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.” —*[3]

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

[4]

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 fleas[5]. 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 Everett[6]. But since this book is also about models of reality, we can question the whole of reality and create a new one that fits.

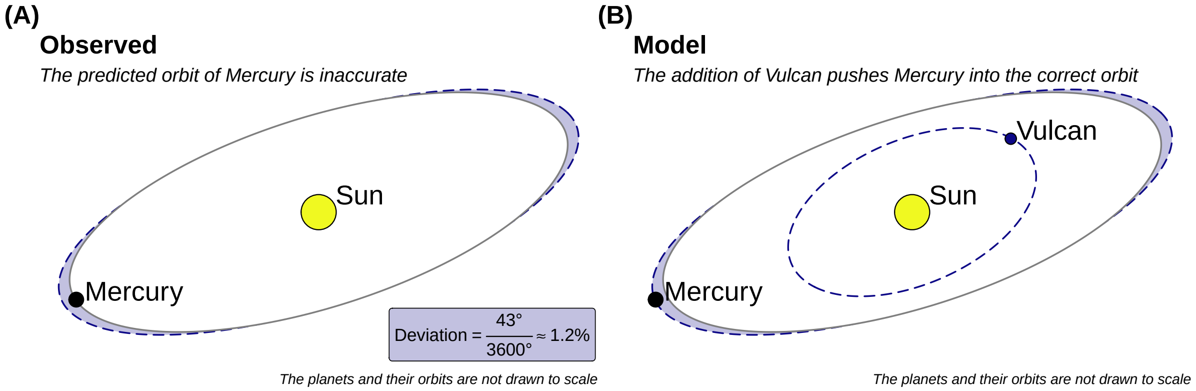
References

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*



Observed

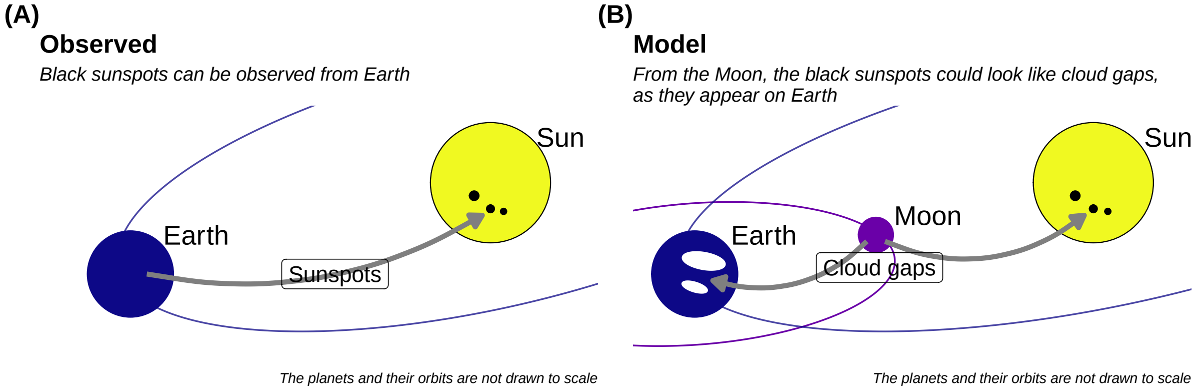
The predicted orbit of Mercury is inaccurate

Model

The addition of Vulcan pushes Mercury into the correct orbit

The planets and their orbits are not drawn to scale

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



Observed

Black sunspots can be observed from Earth

Model

From the Moon, the black sunspots could look like cloud gaps, as they appear on Earth

The planets and their orbits are not drawn to scale

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

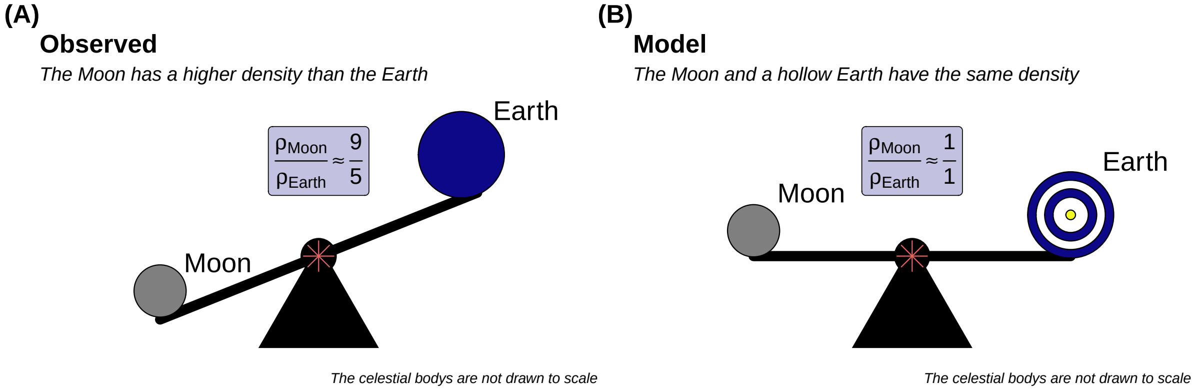

(A)
Observed
The Moon has a higher density than the Earth
$\frac{\rho_{Moon}}{\rho_{Earth}} \approx \frac{9}{5}$
Earth
Moon
The celestial bodys are not drawn to scale
(B)
Model
The Moon and a hollow Earth have the same density
$\frac{\rho_{Moon}}{\rho_{Earth}} \approx \frac{1}{1}$
Moon
Earth
The celestial bodys are not drawn to scale

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

1.1 General background

> *"Any sufficiently advanced technology is indistinguishable from magic."* —[7], *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

[8] What is this thing called Science?

[9] What is science

[10] What is science?

[11] Models Demystified: A Practical Guide from Linear Regression to Deep Learning

[12] Statistical Thinking for the 21st Century

[13] 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

```r
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

$$y = 0.15 \cdot x^3 - 2.2 \cdot x^2 + 8.8 \cdot x + 3.2 \tag{2.1}$$

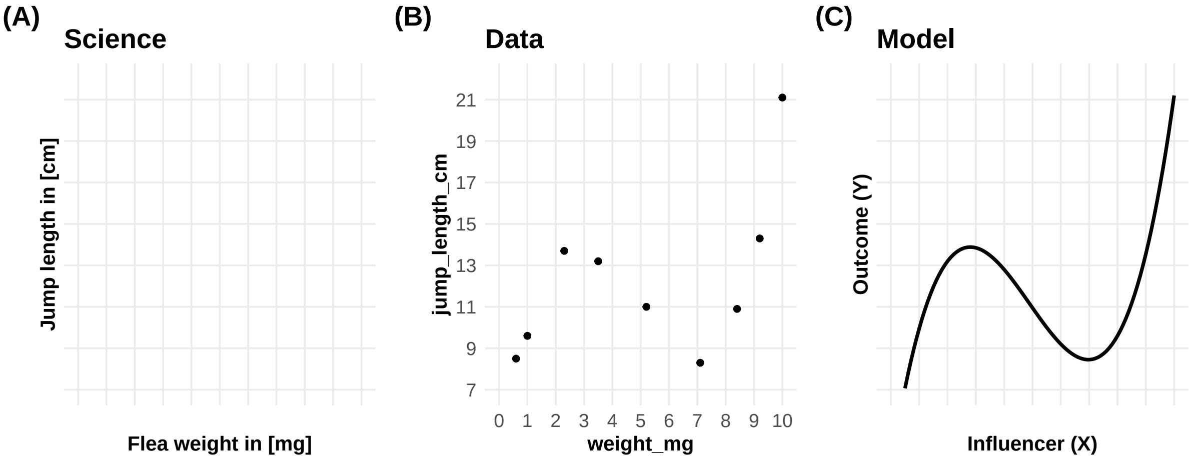


Figure 2.1: foo.

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

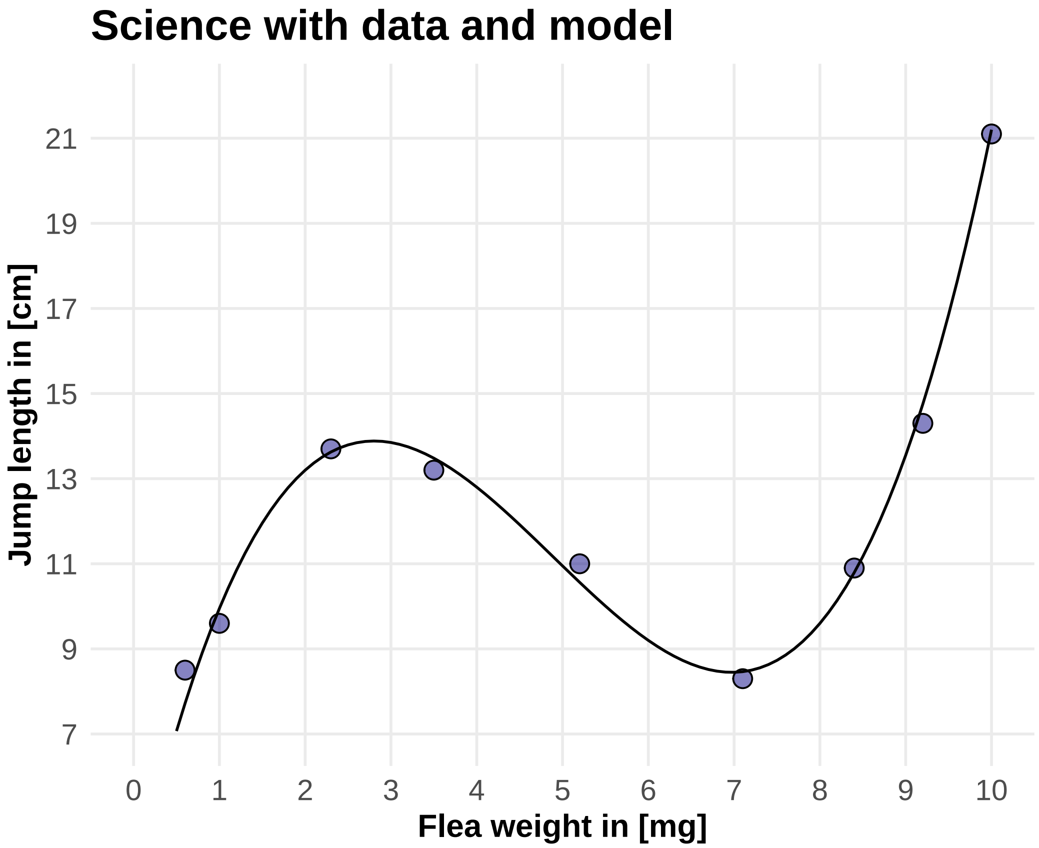


Figure 2.2: foo.

2.8 Summary

References

3 What is a model?

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

Last modified on 19. December 2025 at 19:29:05

> *"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	**outcome**, response, endpoint, dependent variable		*The right-hand side (abbr. RHS) of the model. Describing the values measured in an experiment or study.*

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

Symbol	Name	Application	Description
x	**focal explanator**, **focal predictor**, focal variable	Main effect	*In a model with multiple influencers, the focal variable is the variable of primary interest.*
c	**covariate**, covariable	Continuous x	*The influencer is a numeric variable with continuous values.*
f_A	**factor A**, factorial variable, categorical variable	Categorical x	*The influencer is discrete, functioning as a grouping variable, such as an experimental group or a treatment.*
$A.1$ to $A.j$	**levels**, groups, treatment groups	Factor f_A	*The discrete groups included in one factor A.*

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

[14] What is a statistical model?

[15] What do we mean by a statistical model?

[16] What Is the Purpose of Statistical Modeling?

[17] Where do statistical models come from? Revisiting the problem of specification

[18] Statistical modelling

3.2 Theoretical background

Model notation

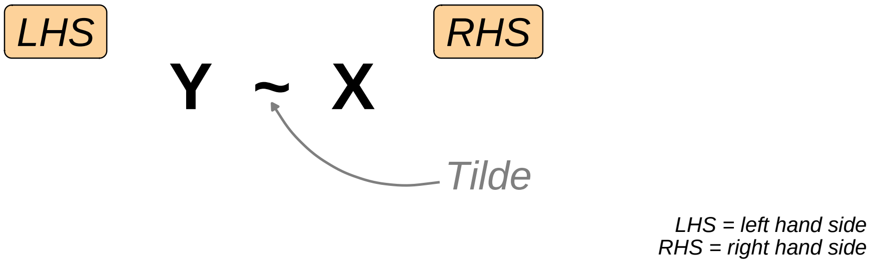


Figure 3.1: foo

Data, model, and error

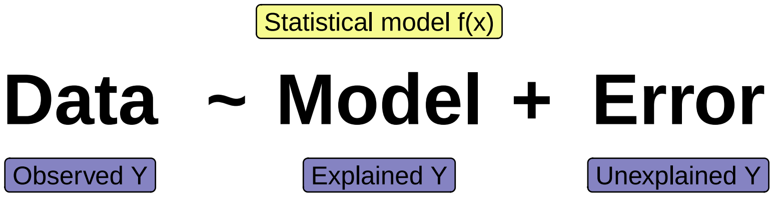


Figure 3.2: foo

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

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

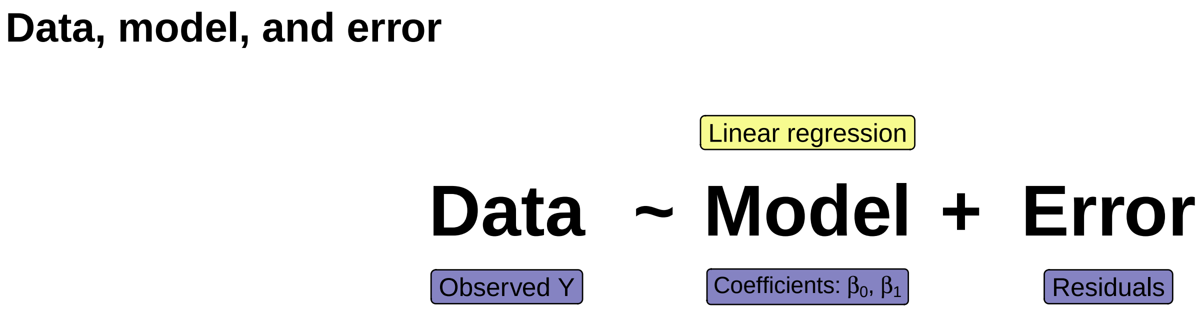


Figure 3.3: foo

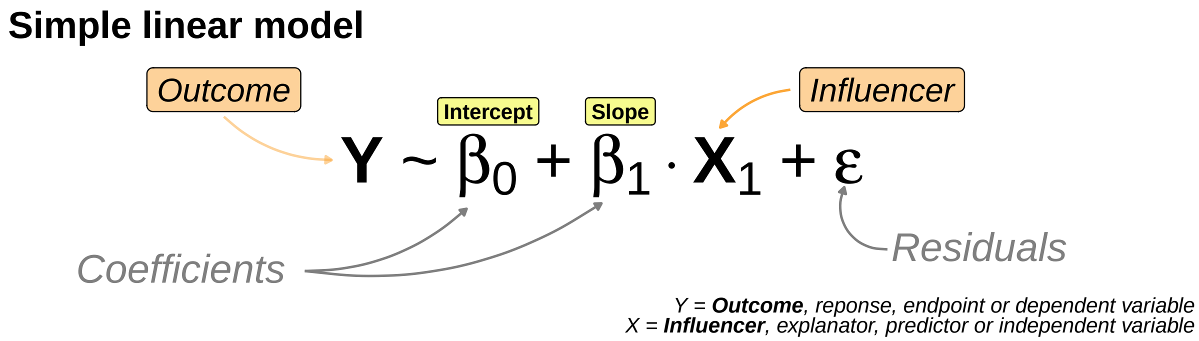

Simple linear model
Outcome
Intercept
Slope
Influencer
$Y \sim \beta_0 + \beta_1 \cdot X_1 + \varepsilon$
Coefficients
Residuals
Y = Outcome, reponse, endpoint or dependent variable
X = Influencer, explanator, predictor or independent variable

Figure 3.4: foo

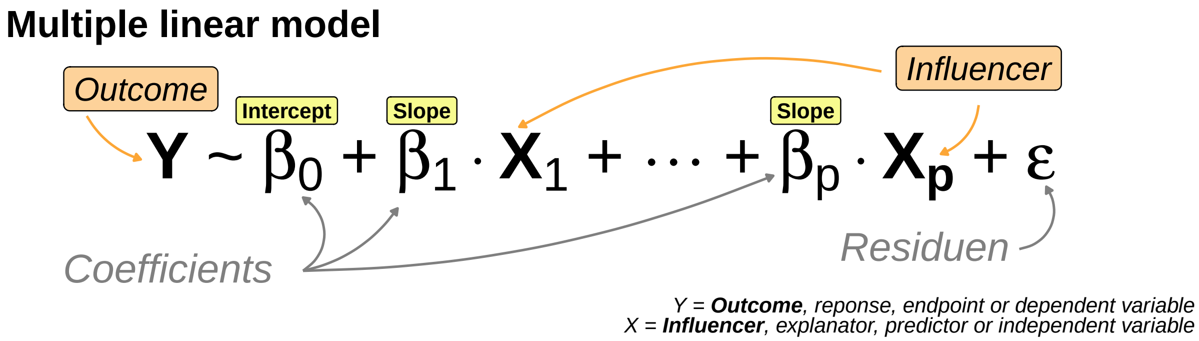

Multiple linear model
Outcome
Intercept
Slope
Slope
Influencer
$Y \sim \beta_0 + \beta_1 \cdot X_1 + \cdots + \beta_p \cdot X_p + \varepsilon$
Coefficients
Residuen
Y = Outcome, reponse, endpoint or dependent variable
X = Influencer, explanator, predictor or independent variable

Figure 3.5: foo

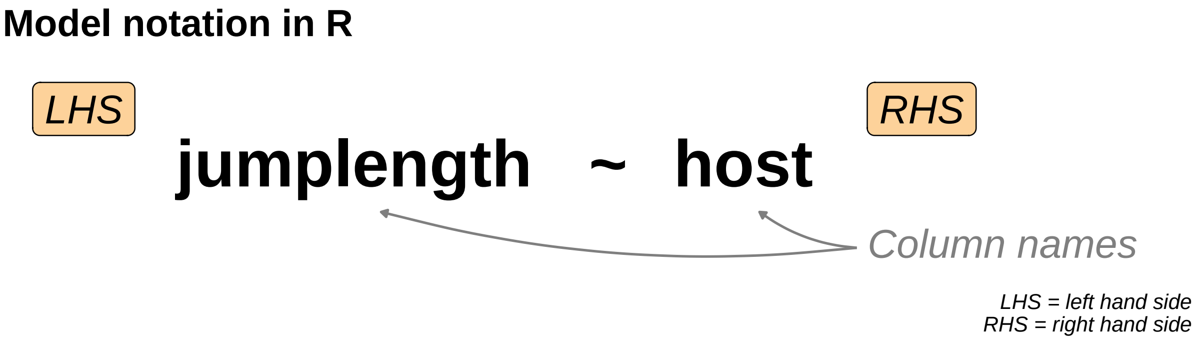


Figure 3.6: foo

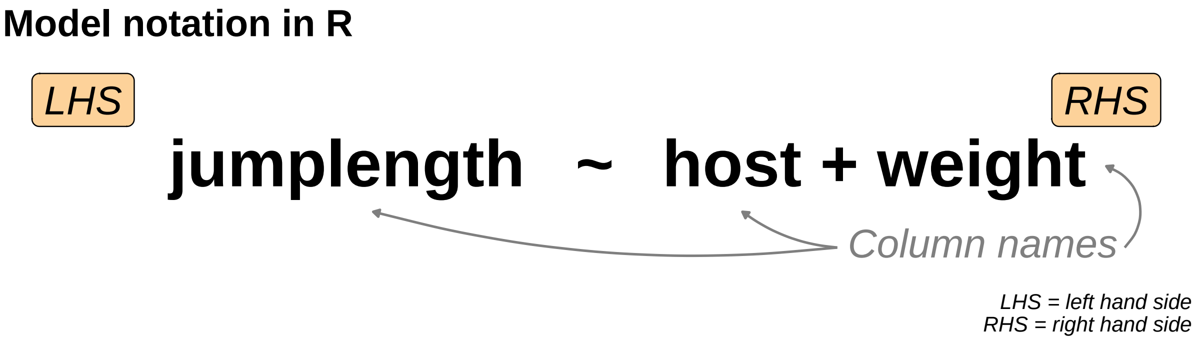


Figure 3.7: foo

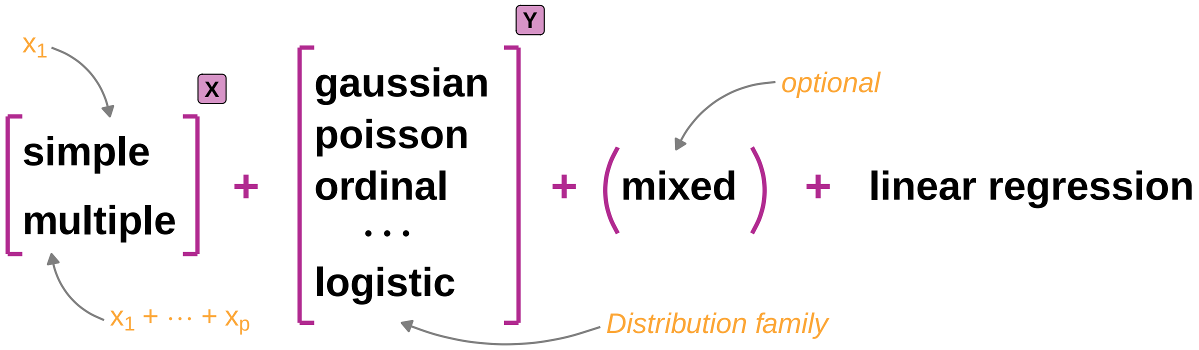


Figure 3.8: foo

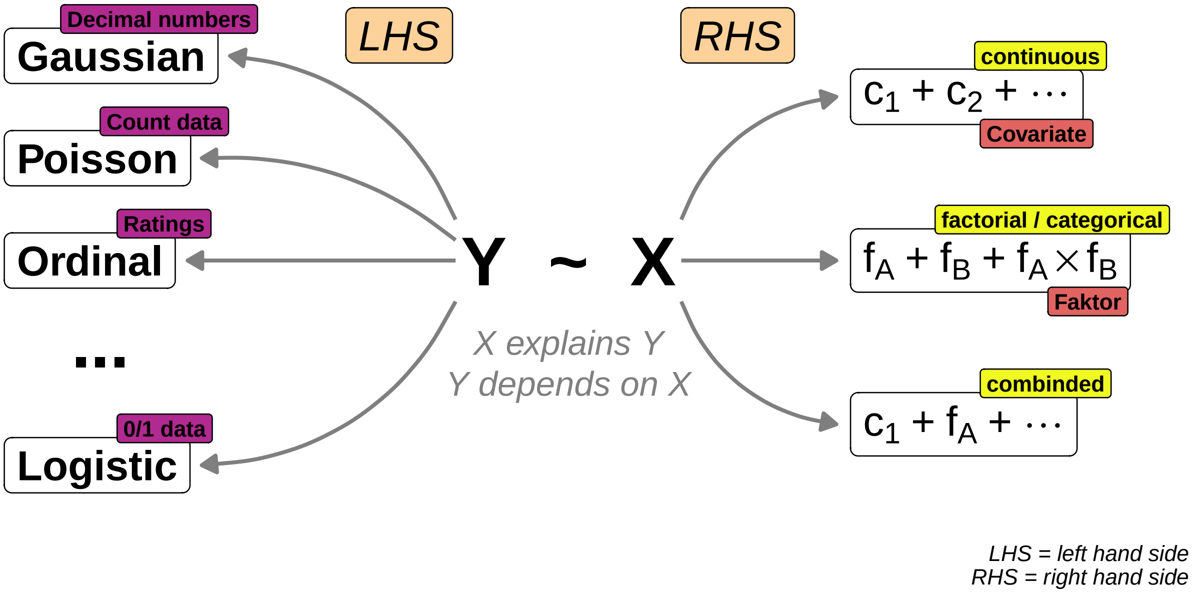


Figure 3.9: foo

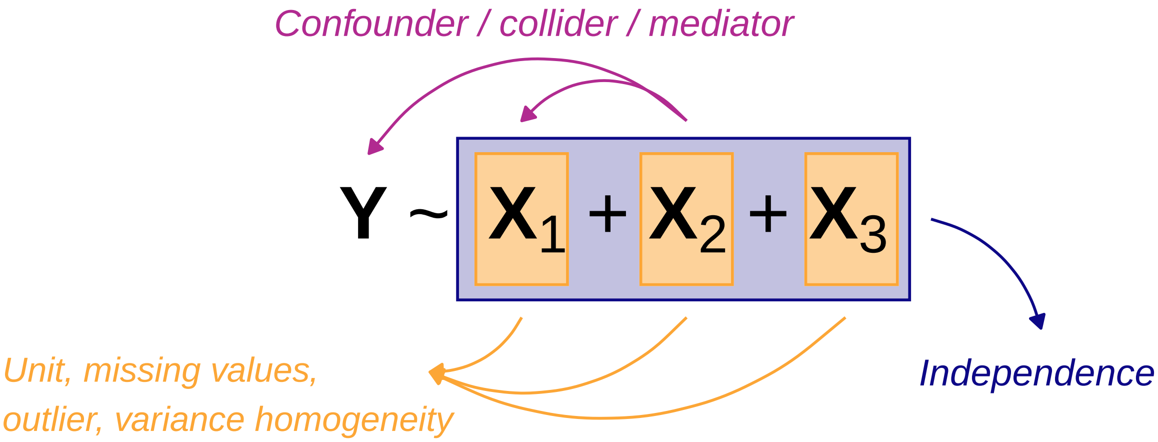

Confounder / collider / mediator
$Y \sim X_1 + X_2 + X_3$
Unit, missing values,
outlier, variance homogeneity
Independence

Figure 3.10: foo

> *"Twistin', shake it, shake it, shake it, shake it, baby; Hey, we gonna Loop de Loop Shake it out, baby; Hey, we gonna Loop de La"* — *The Blues Brothers (1980) – Shake a Tail Feather*

> *"It started with a ~~kiss~~ line; [...] How could I resist?"* — *Hot Chocolate – It Started With A Kiss*

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 09. December 2025 at 10:28:31

> *“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 fleas[5]. 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 branches[6].

References

4 Body weight of fleas

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

Last modified on 22. December 2025 at 19:58:29

> *"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)
```

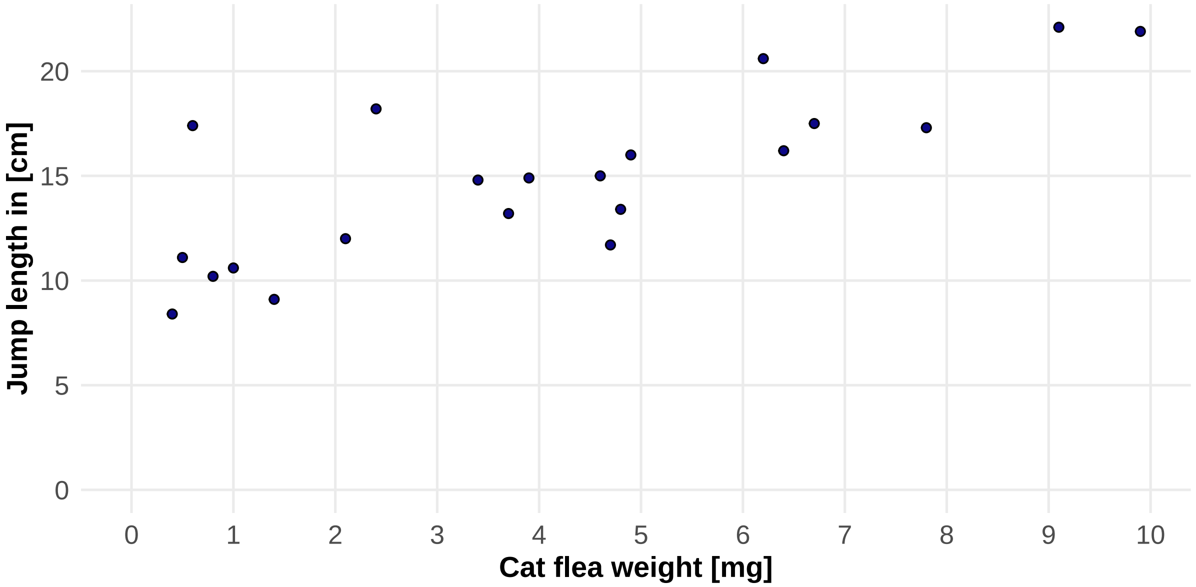


Figure 4.1: foo **(A)** foo **(B)** foo

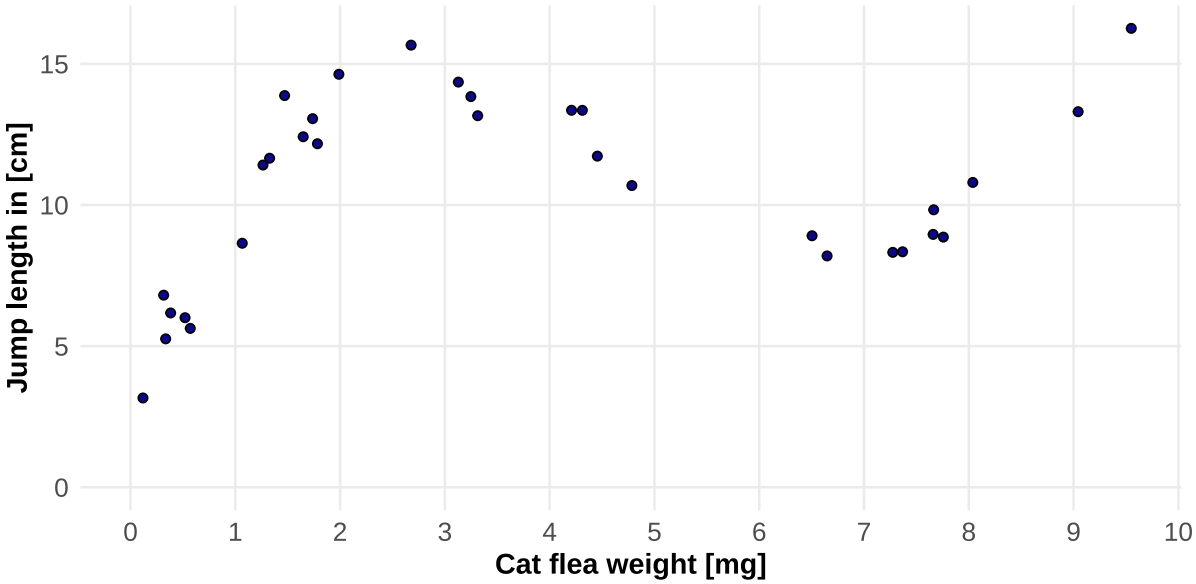


Figure 4.2: foo **(A)** foo **(B)** foo

4.5 Data availability

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

4.6 Alternatives

Further tutorials and R packages on XXX

4.7 Glossary

term what does it mean.

4.8 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

4.9 Summary

References

5 Fleas of animals

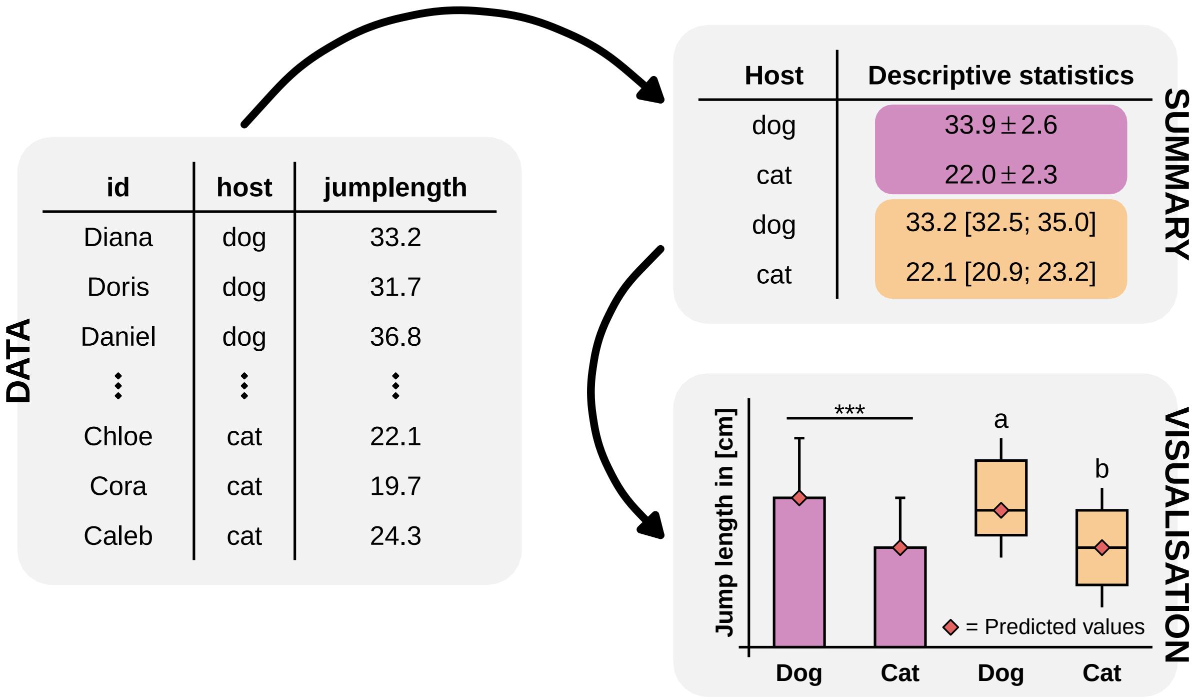
Last modified on 14. December 2025 at 19:22:00

> *"A quote."* — *Dan Meyer*

5.1 General background

What is a data grid?

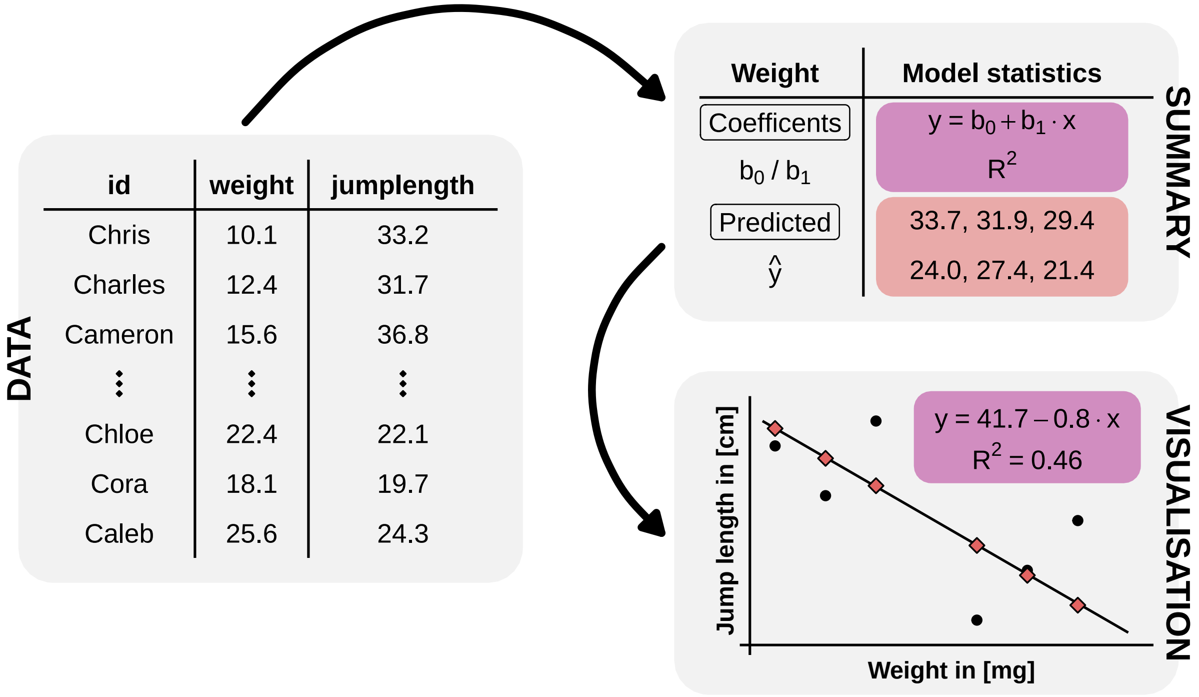
5.2 Theoretical background



id	host	jumplength
Diana	dog	33.2
Doris	dog	31.7
Daniel	dog	36.8
⋮	⋮	⋮
Chloe	cat	22.1
Cora	cat	19.7
Caleb	cat	24.3

Host	Descriptive statistics
dog	33.9 ± 2.6
cat	22.0 ± 2.3
dog	33.2 [32.5; 35.0]
cat	22.1 [20.9; 23.2]

Figure 5.1: foo



id	weight	jumplength
Chris	10.1	33.2
Charles	12.4	31.7
Cameron	15.6	36.8
⋮	⋮	⋮
Chloe	22.4	22.1
Cora	18.1	19.7
Caleb	25.6	24.3

Weight	Model statistics
Coefficents b_0 / b_1	$y = b_0 + b_1 \cdot x$ R^2
Predicted $\hat{y}$	33.7, 31.9, 29.4 24.0, 27.4, 21.4

Figure 5.2: foo

5.3 R packages used

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

19

5.4 Data

Small data and grid

5.4.1 Jump performances of fleas

5

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

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

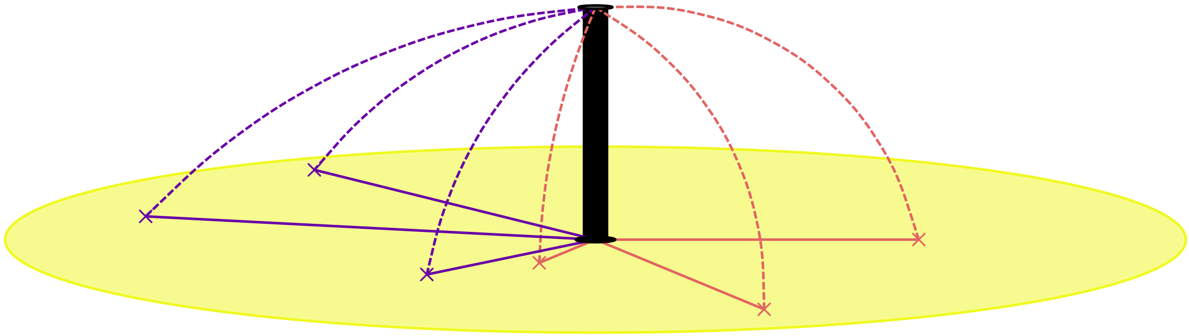


Figure 5.3: foo **(A)** foo **(B)** foo

```
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))
```

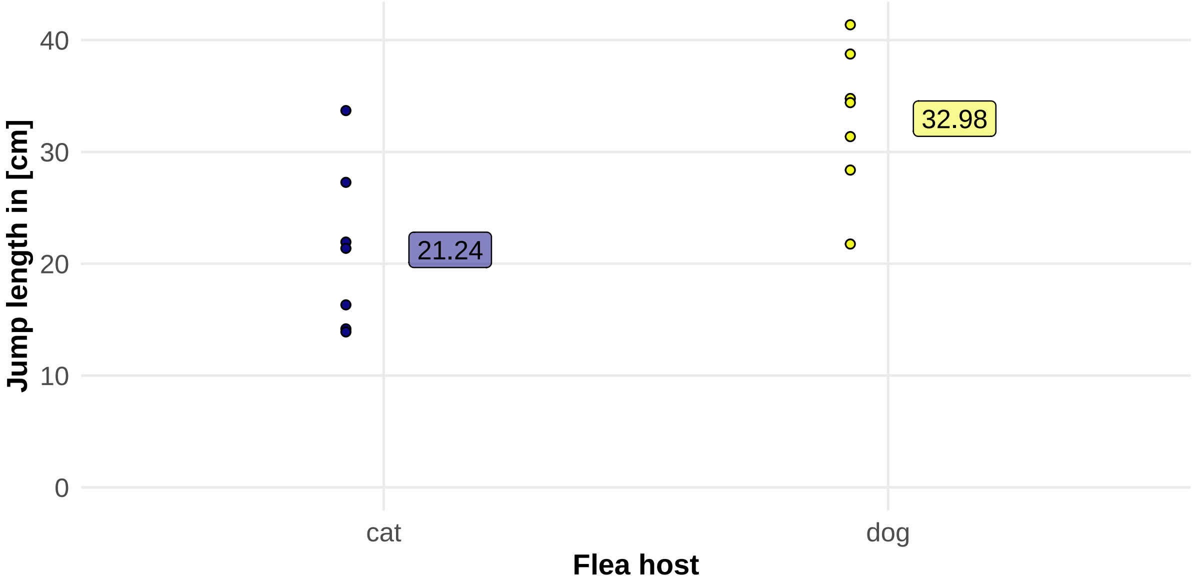


Figure 5.4: foo **(A)** foo **(B)** foo

```
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
```

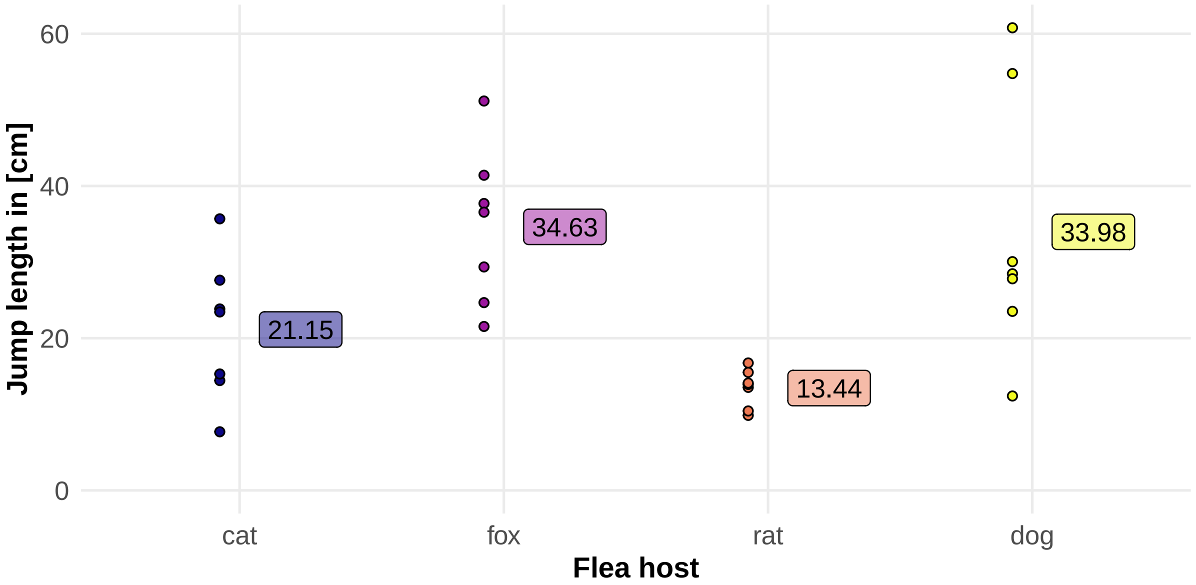


Figure 5.5: foo **(A)** foo **(B)** foo

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),
       counthairleg_left = rpois(7, 4),
       counthairleg_right = rpois(7, 4),
       counthairleg = (counthairleg_left + counthairleg_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 counthairleg_left counthairleg_right counthairleg 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?

20

```
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
```

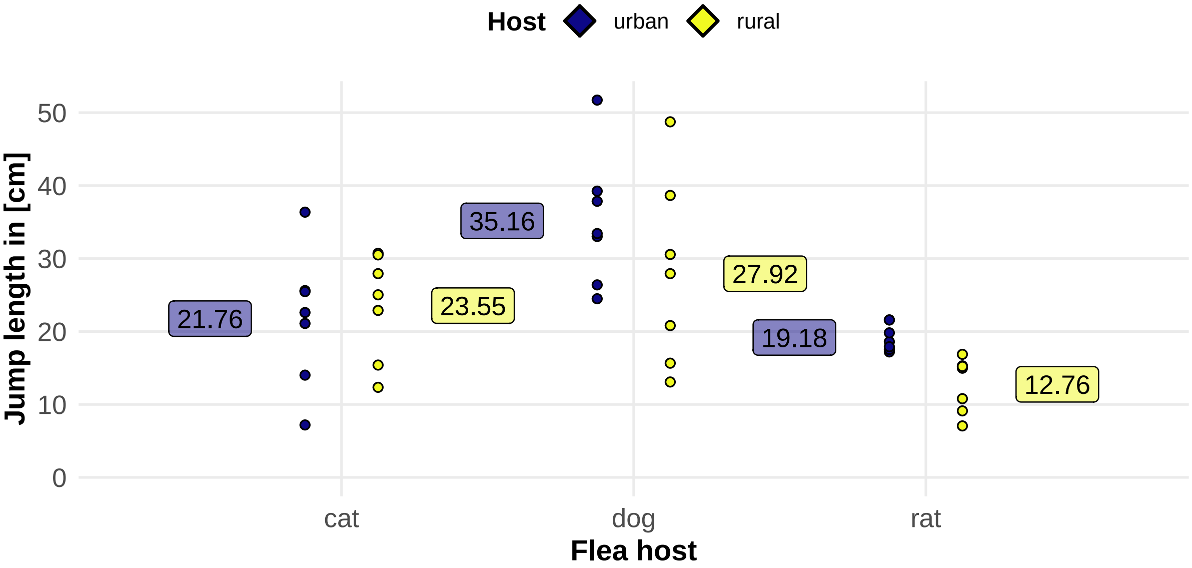


Figure 5.6: foo **(A)** foo **(B)** foo

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)
```

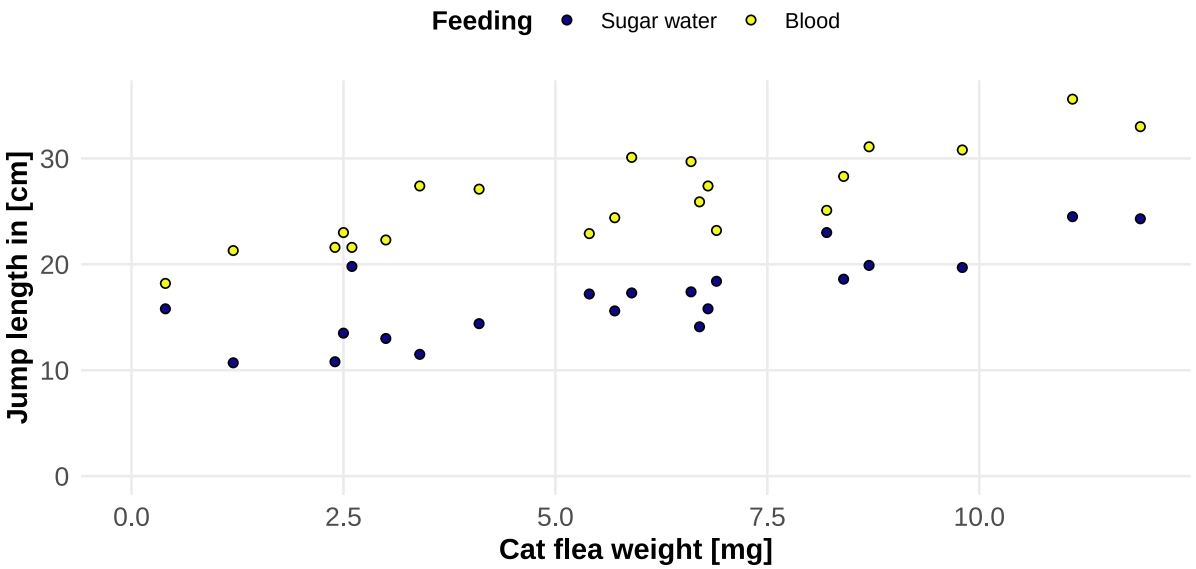


Figure 6.1: foo **(A)** foo **(B)** foo

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)
```

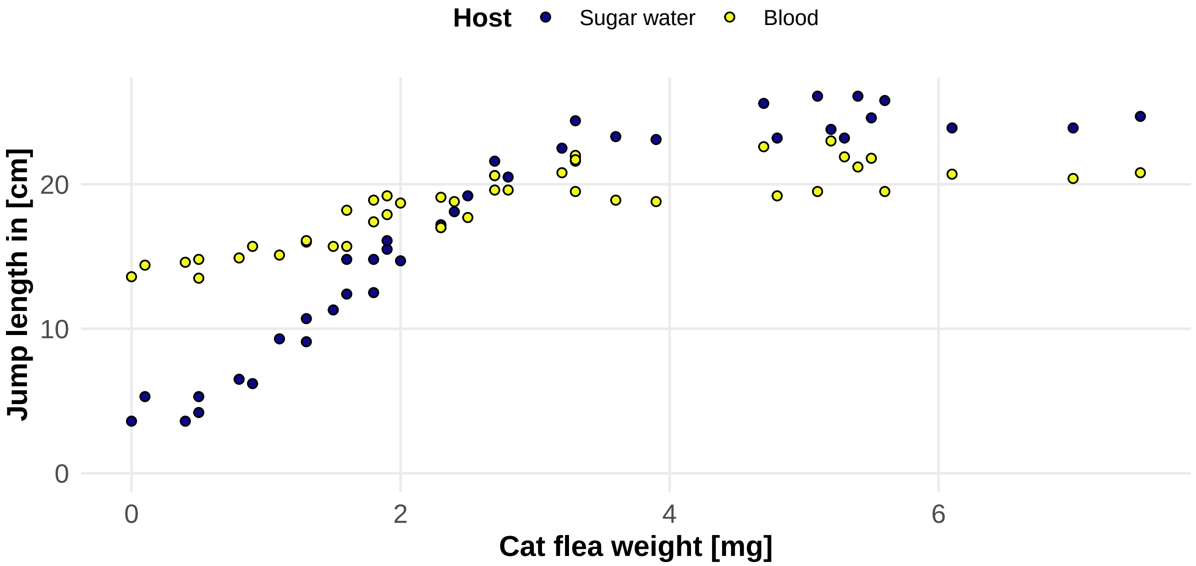


Figure 6.2: foo **(A)** foo **(B)** foo

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
```

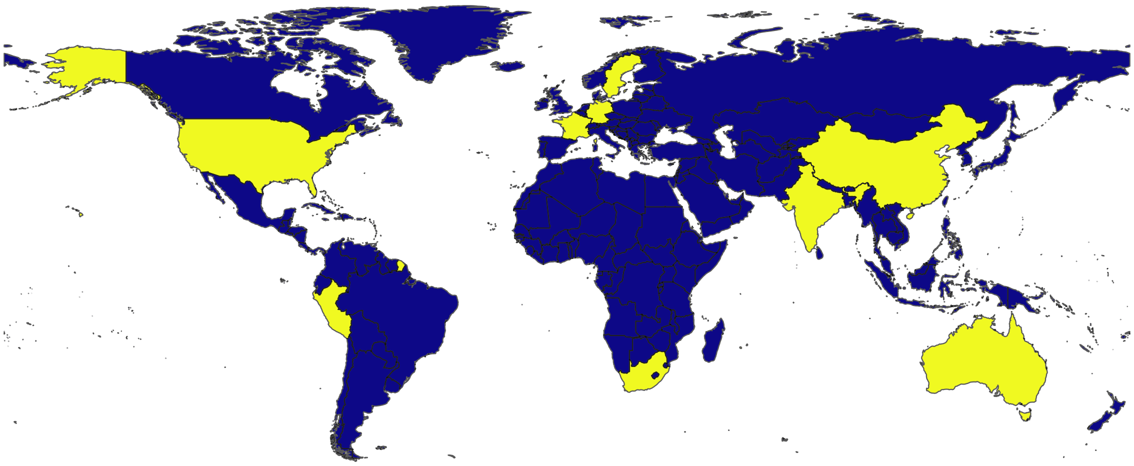


Figure 6.3: foo **(A)** foo **(B)** foo

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)

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)     5.320
          sitesemi-dense 4.307    -0.34
          sitecity        4.251    -0.35  0.36
 Residual                 2.926
```

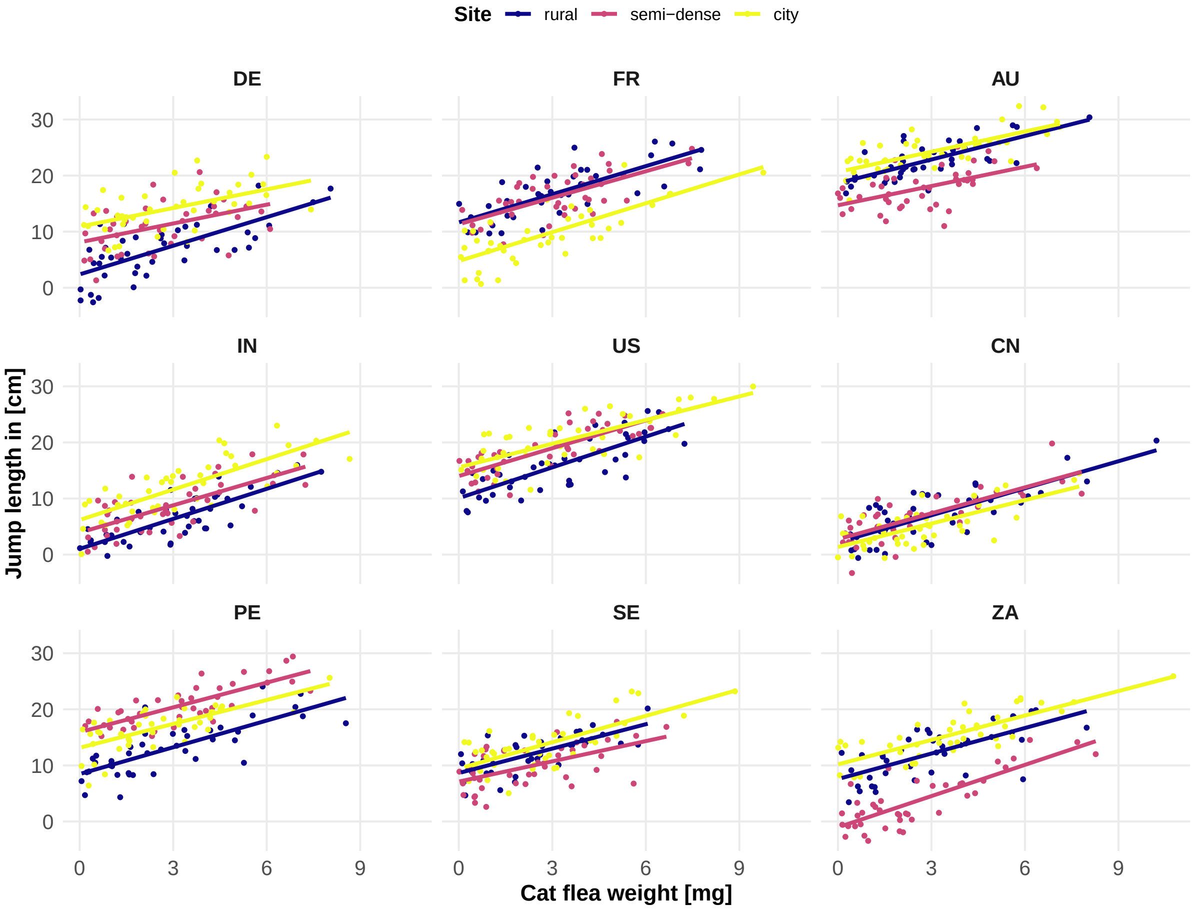


Figure 6.4: foo **(A)** foo **(B)** foo

```
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 09. December 2025 at 10:28:21

> *"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

21

19

22

23

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

Hypothese testing

Last modified on 09. December 2025 at 10:28:06

> *"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 11. December 2025 at 20:23:05

> *"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.

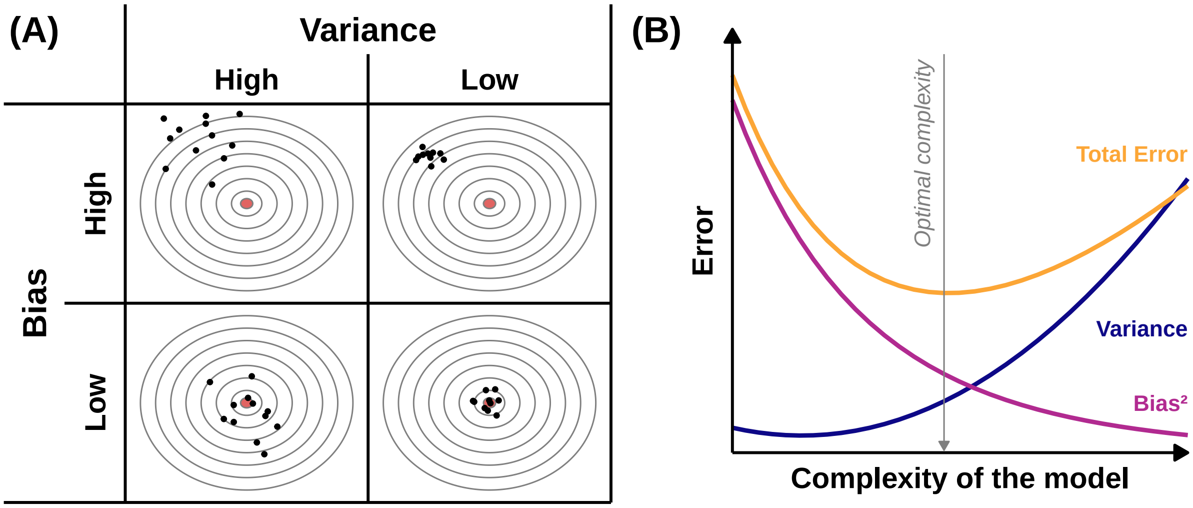
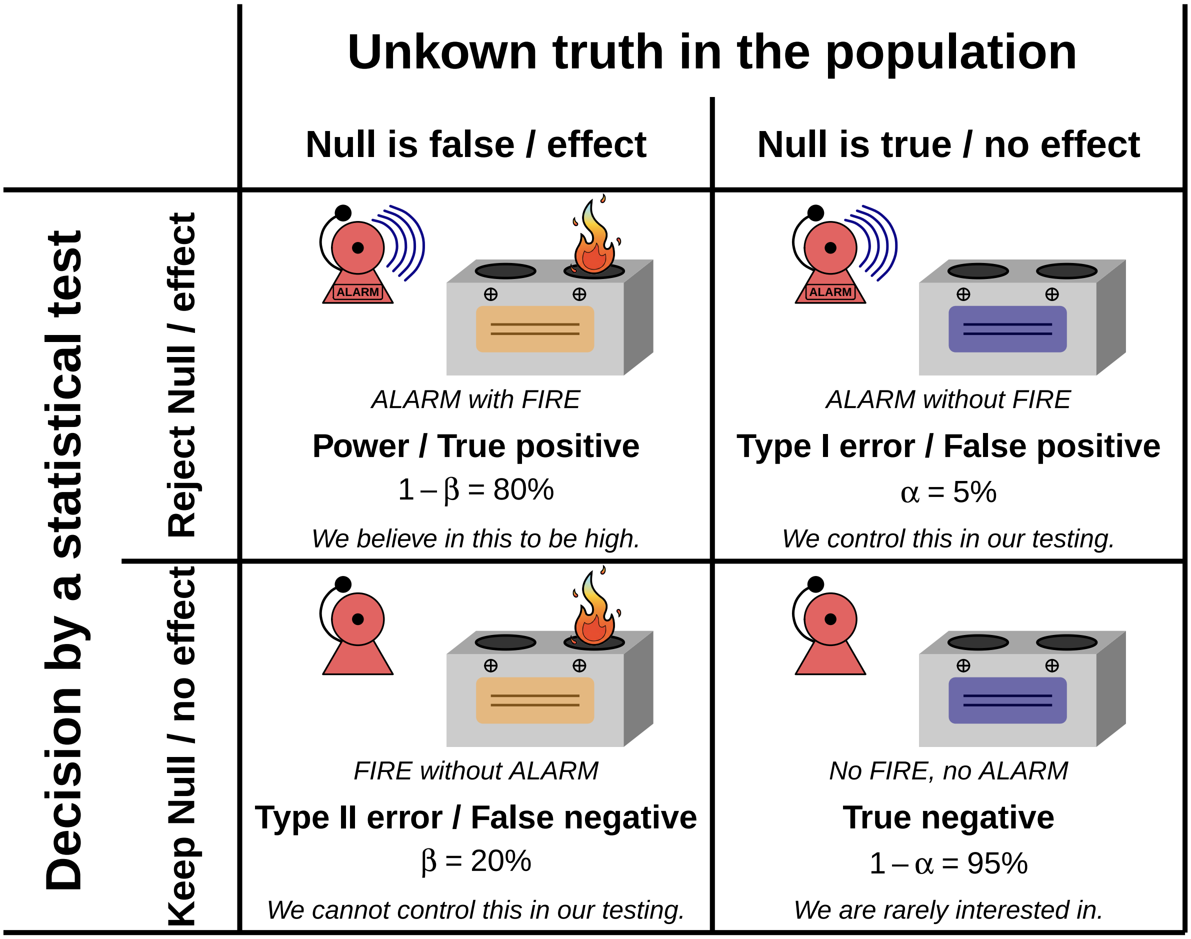


Figure 8.1: foo



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

Figure 8.2: foo

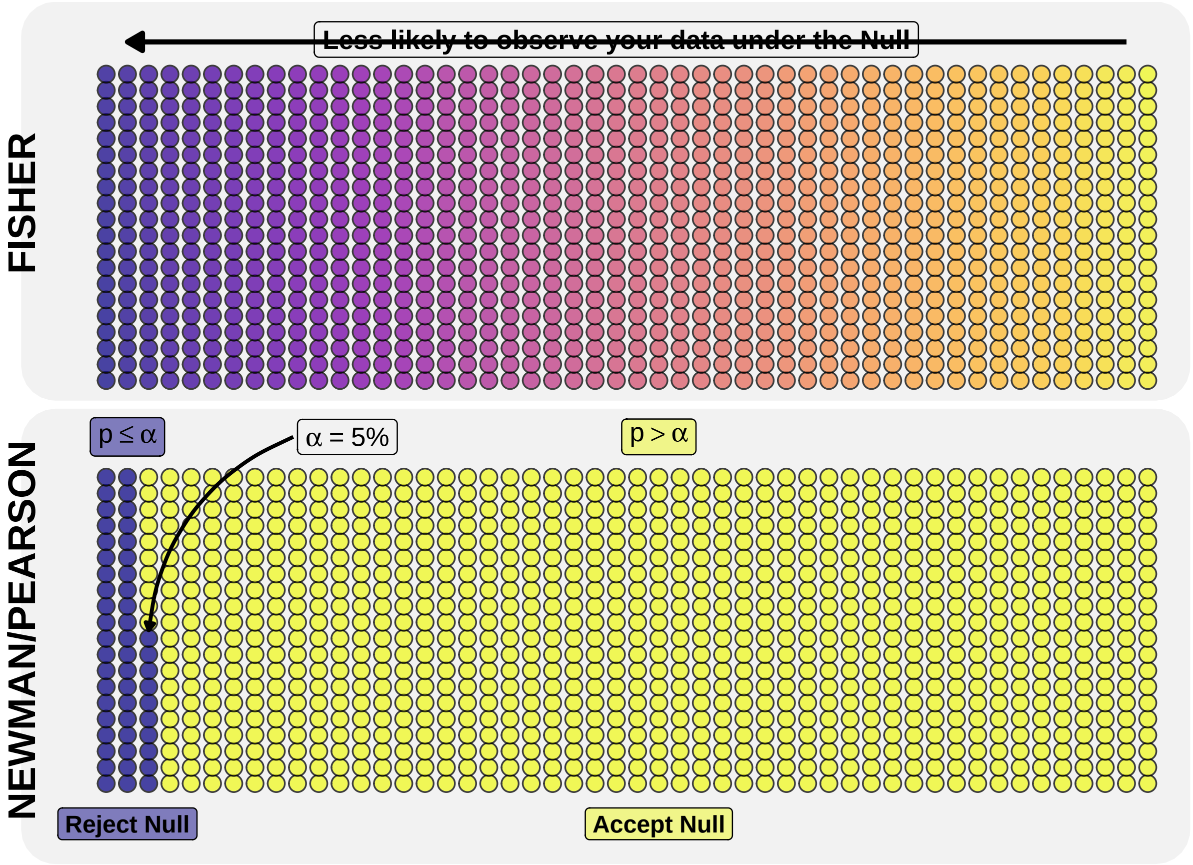


Figure 8.3: foo

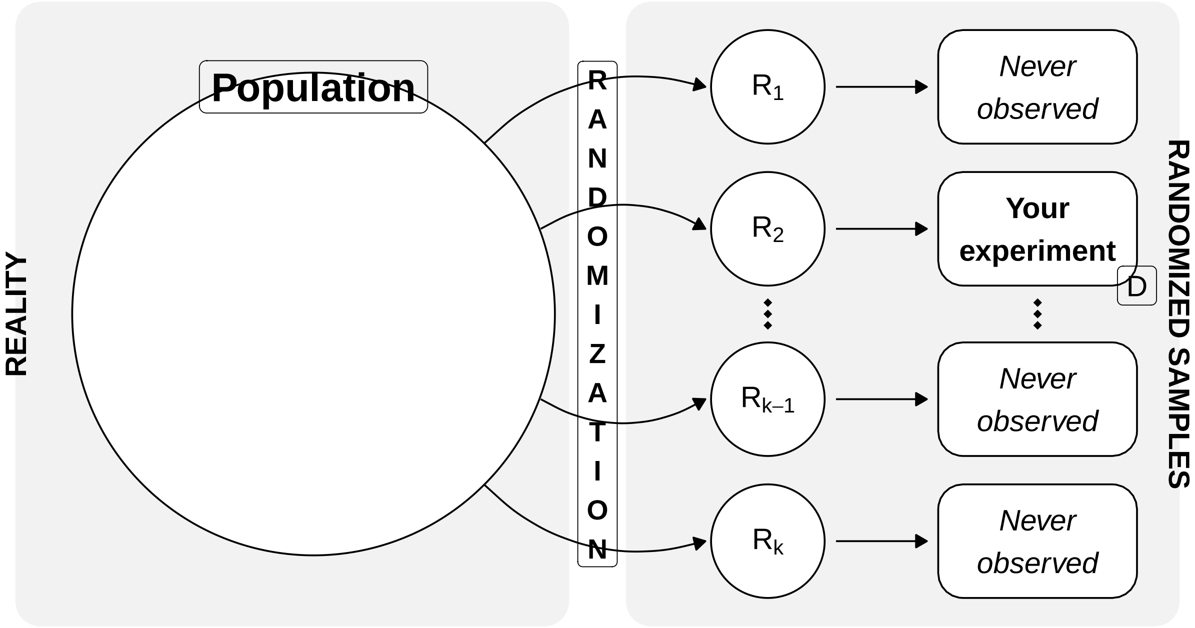


Figure 8.4: foo

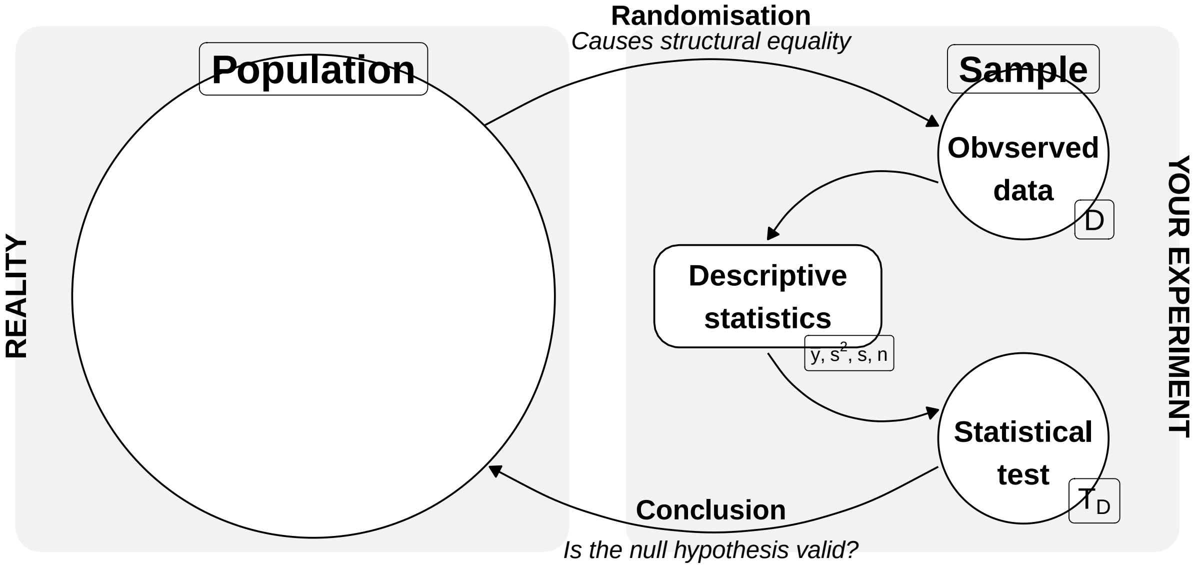


Figure 8.5: foo

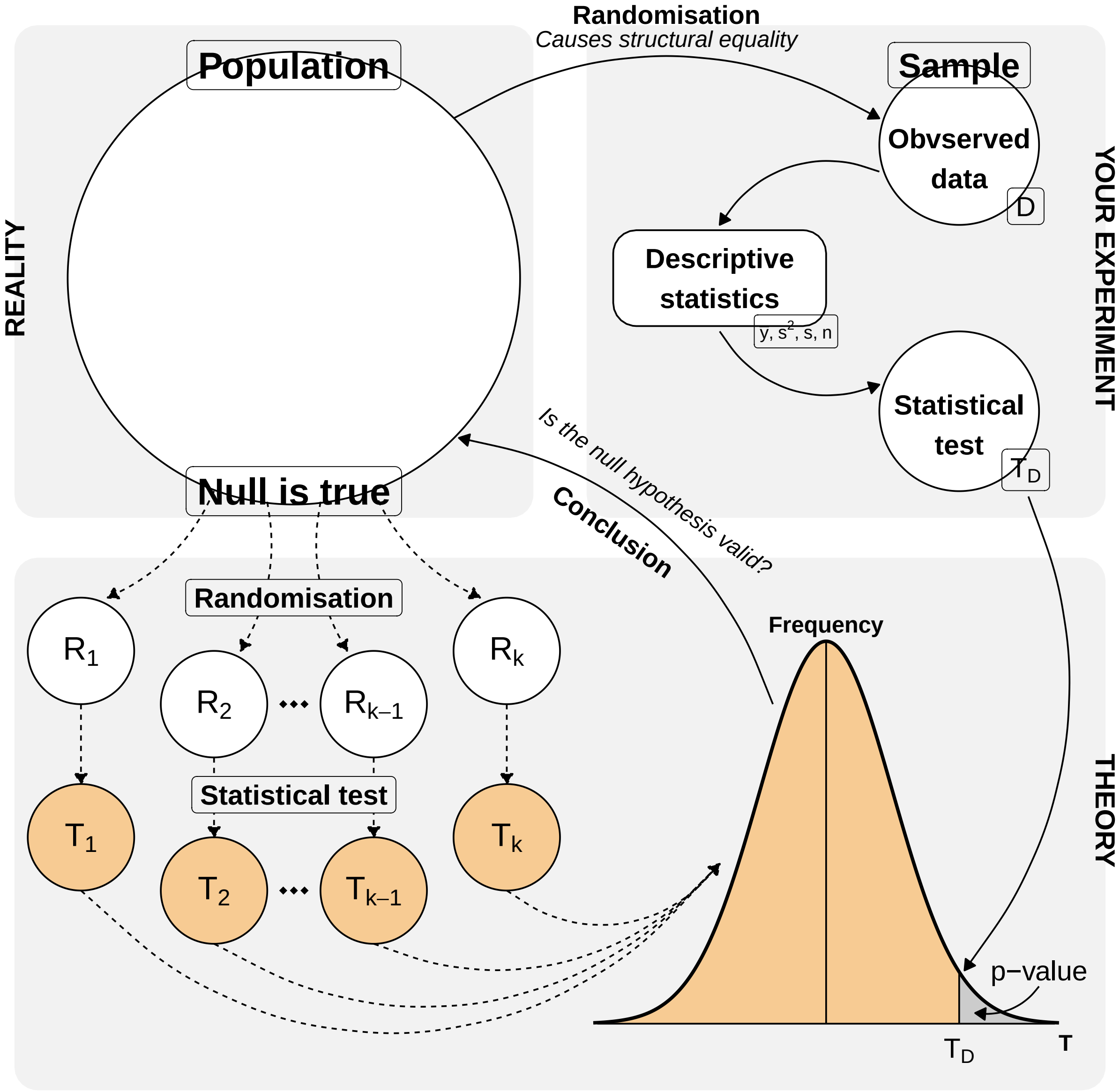


Figure 8.6: foo

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).

- 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

Visualisation of data

Last modified on 20. December 2025 at 20:24:43

> *"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

```r
pacman::p_load(emmeans, parameters, nlme, broom)

foo <- tibble(A = rnorm(10000, 5, 4),
              B = rnorm(10000, 8, 2)) |>
  gather()

foo |>
  group_by(key) |>
  summarise(mean(value), var(value), sd(value))
```

```
# A tibble: 2 x 4
  key   `mean(value)` `var(value)` `sd(value)`
  <chr>         <dbl>        <dbl>       <dbl>
1 A              4.99        16.1         4.02
2 B              8.01         4.03        2.01
```

```r
sqrt((16.77840 + 3.94372)/2)
```

```
[1] 3.21886
```

```r
fit <- lm(value ~ 0+key, foo)

fit |> parameters()
```

```
Parameter | Coefficient |   SE |       95% CI | t(19998) |      p
-----------------------------------------------------------------
key [A]   |        4.99 | 0.03 | [4.92, 5.05] |   157.01 | < .001
key [B]   |        8.01 | 0.03 | [7.95, 8.07] |   252.19 | < .001


Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
  using a Wald t-distribution approximation.
```

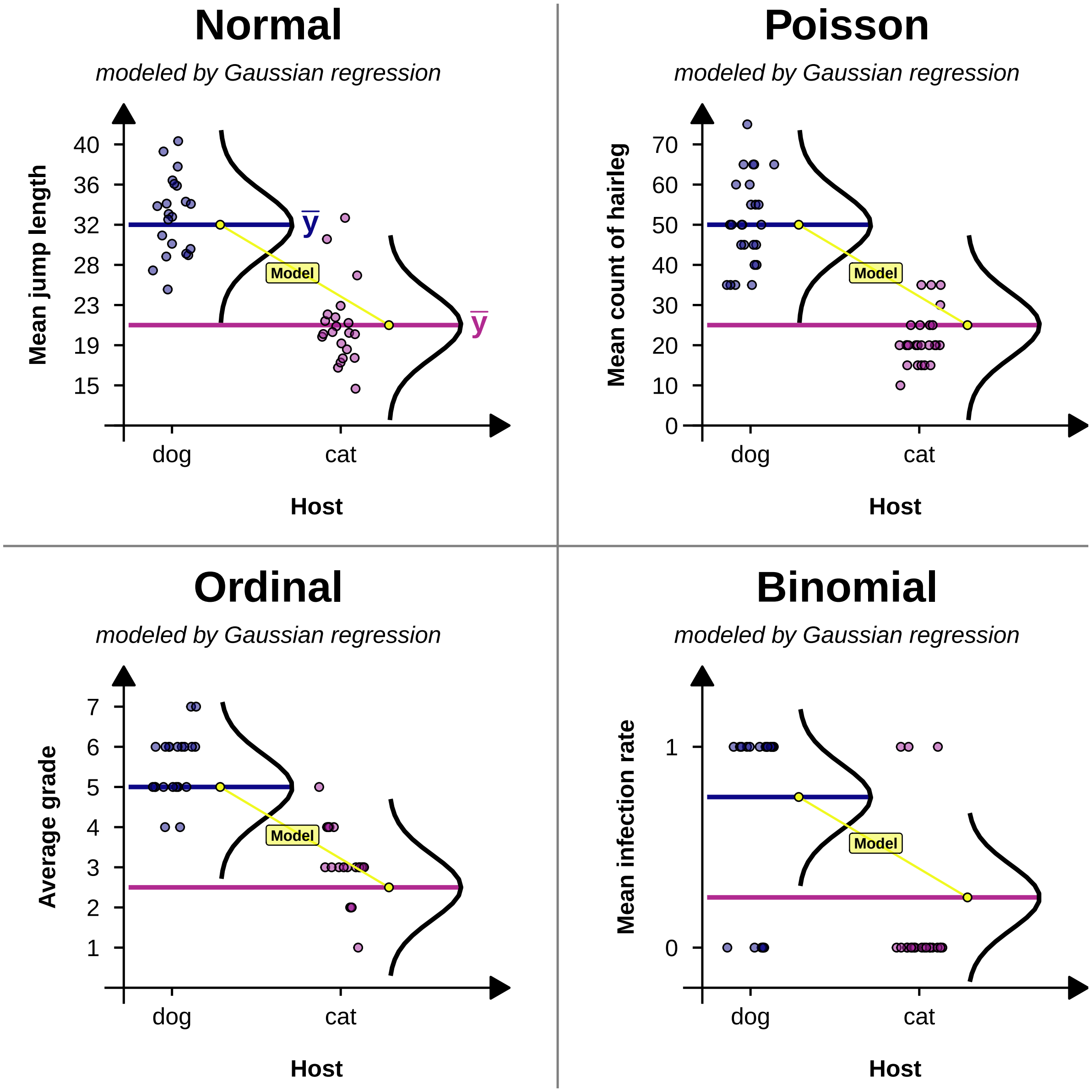


Figure 8.7: foo

```r
fit |> glance()
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df  logLik     AIC     BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl>
1     0.815         0.815  3.18    44126.       0     2 -51486. 102979. 103002.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```r
0.02012848 * sqrt(10000)
```

```
[1] 2.012848
```

```r
sqrt(diag(vcov(fit)))
```

```
      keyA       keyB 
0.03175371 0.03175371 
```

```r
model_parameters(fit, vcov = "HC3")
```

```
Parameter | Coefficient |   SE |       95% CI | t(19998) |      p
-----------------------------------------------------------------
key [A]   |        4.99 | 0.04 | [4.91, 5.06] |   124.10 | < .001
key [B]   |        8.01 | 0.02 | [7.97, 8.05] |   399.10 | < .001


Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
  using a Wald t-distribution approximation.
```

```r
gls(value ~ 0 + key, weights = varIdent(form =  ~ 1 | key), foo) |>
  parameters()
```

```
# Fixed Effects

Parameter | Coefficient |   SE |       95% CI | t(19998) |      p
-----------------------------------------------------------------
key [A]   |        4.99 | 0.04 | [4.91, 5.06] |   124.10 | < .001
key [B]   |        8.01 | 0.02 | [7.97, 8.05] |   399.12 | < .001
```

```
Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
  using a Wald t-distribution approximation.
```

```r
emm <- emmeans(fit, "key", vcov = sandwich::vcovHAC)
summary_emm <- summary(emm)
# Calculate SD from SE and sample size (n)
summary_emm$SE * sqrt(summary_emm$df/2)
```

```
[1] 4.015578 2.002160
```

9 Explorative data analysis

Last modified on 26. December 2025 at 19:09:05

> *"A quote." — Dan Meyer*

9.1 General background

9.2 Theoretical background

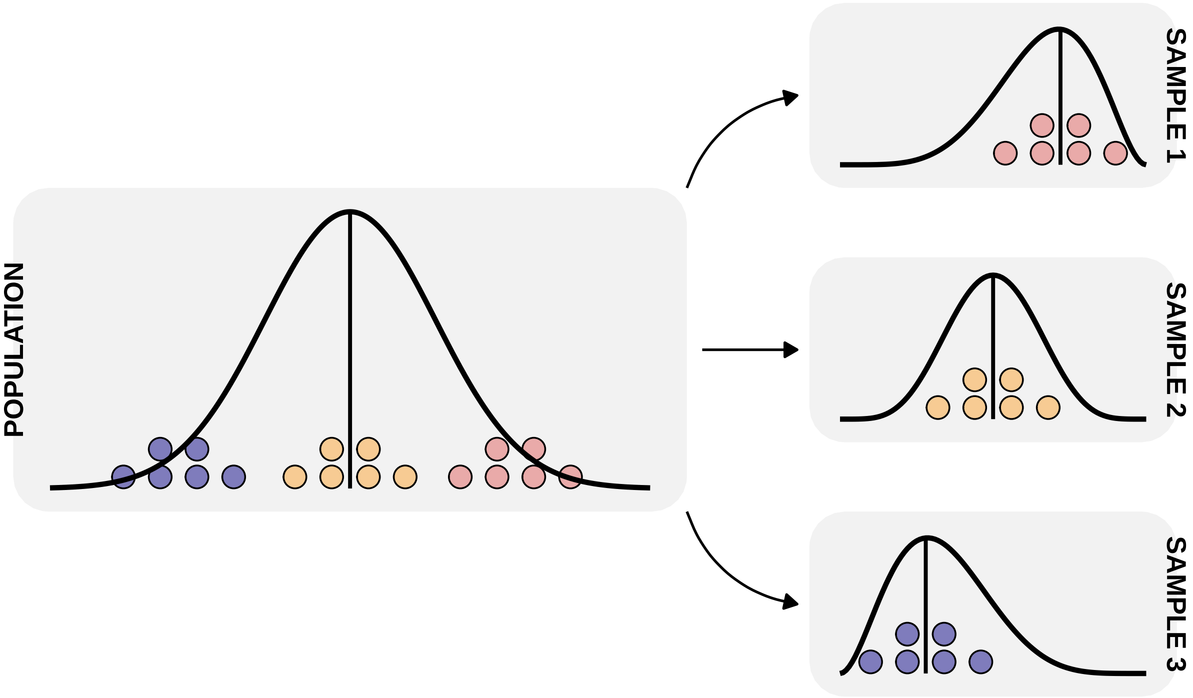


Figure 9.1: foo

9.3 R packages used

9.4 Data

```r
jump_dog_tbl <- read_delim("data/jump_fleas.txt") |>
  select(host, jumplength) |>
  filter(host == "dog")
```

```r
jump_dog_tbl
```

```
# A tibble: 7 x 2
  host  jumplength
  <chr>      <dbl>
1 dog         21.8
2 dog         28.4
3 dog         34.8
4 dog         34.4
5 dog         38.8
6 dog         31.4
7 dog         41.4
```

```r
sem <- \(x) sd(x)/sqrt(length(x))

sem(jump_dog_tbl$jumplength)
```

```
[1] 2.480887
```

```r
sem <- \(x) sqrt(var(x)/length(x))

sem(jump_dog_tbl$jumplength)
```

```
[1] 2.480887
```

```r
sem <- \(x) sqrt(sum((x-mean(x))^2)/(length(x)*(length(x)-1)))

sem(jump_dog_tbl$jumplength)
```

```
[1] 2.480887
```

```r
sem <- \(x) sqrt(sum((x-mean(x))^2))/sqrt(length(x)*(length(x)-1))

sem(jump_dog_tbl$jumplength)
```

```
[1] 2.480887
```

$$SE = \frac{s}{\sqrt{n}}$$

$$SE = \sqrt{\frac{s^2}{n}}$$

$$SE = \frac{\sqrt{SS}}{n}; \text{ for large n}$$

$$SE = \sqrt{\frac{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n \cdot (n-1)}} = \frac{\sqrt{SS}}{\sqrt{n \cdot (n-1)}}$$

With n is large

$$SE = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{n^2}} = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{n} = \frac{\sqrt{SS}}{n}$$

```r
jump_dog_tbl |>
  group_by(host) |>
  summarise(mean(jumplength), var(jumplength), sd(jumplength), sem(jumplength)) |>
  mutate_if(is.numeric, round, 2)
```

```
# A tibble: 1 x 5
  host  `mean(jumplength)` `var(jumplength)` `sd(jumplength)` `sem(jumplength)`
  <chr>              <dbl>             <dbl>            <dbl>             <dbl>
1 dog                 33.0              43.1             6.56              2.48
```

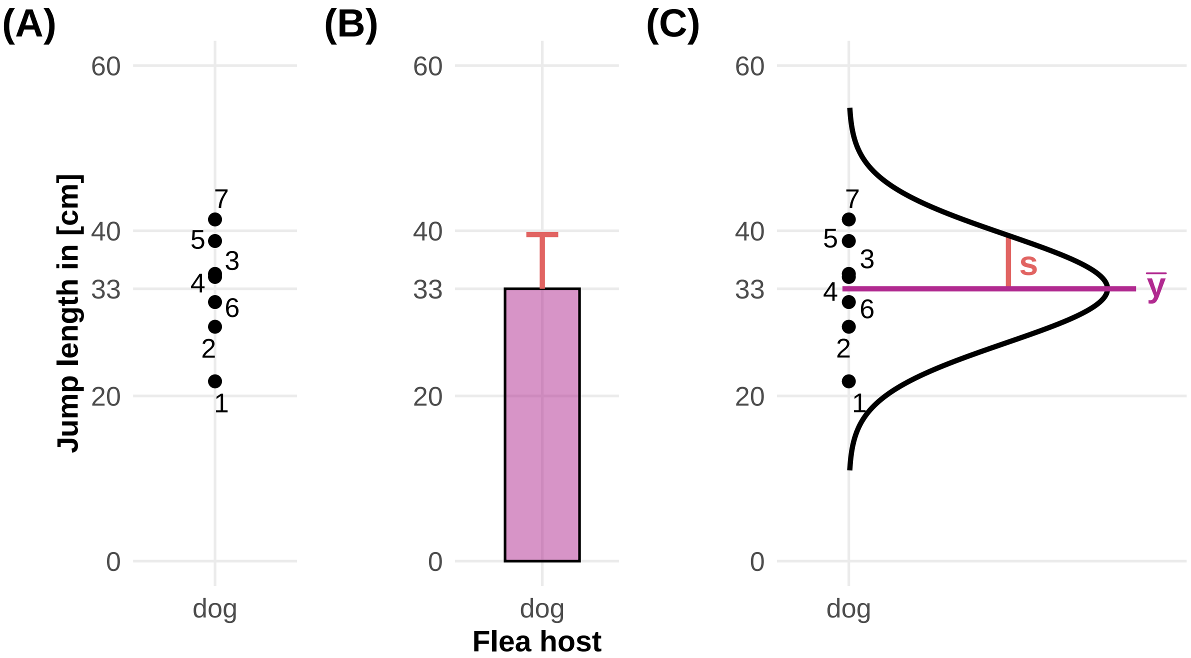


Figure 9.2: foo

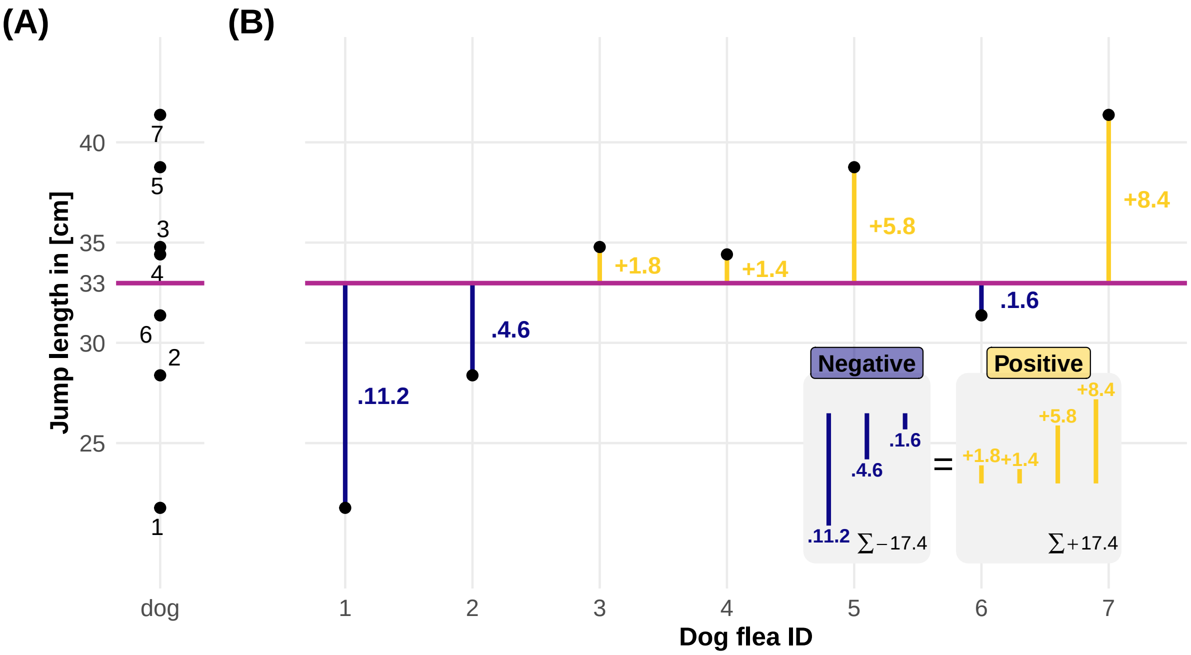


Figure 9.3: foo

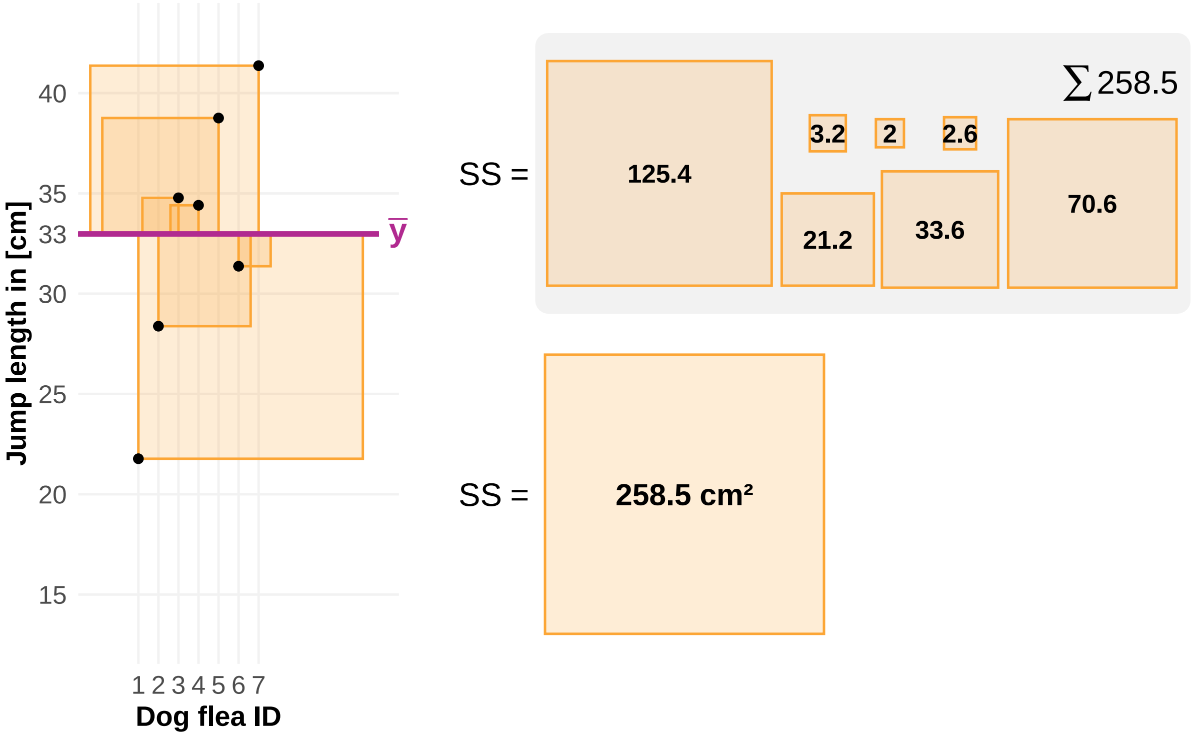


Figure 9.4: foo

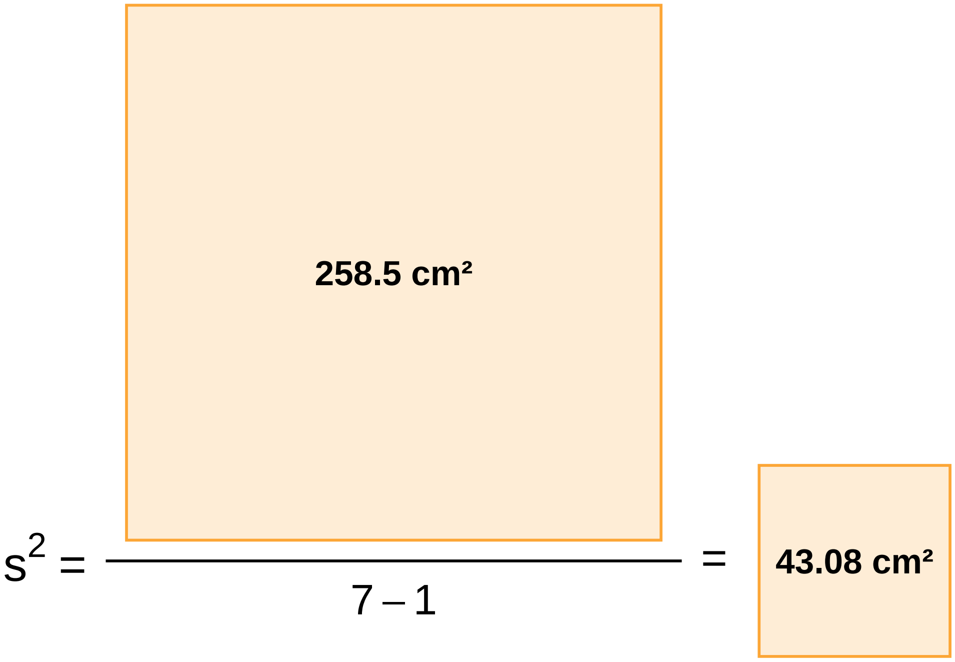


Figure 9.5: foo

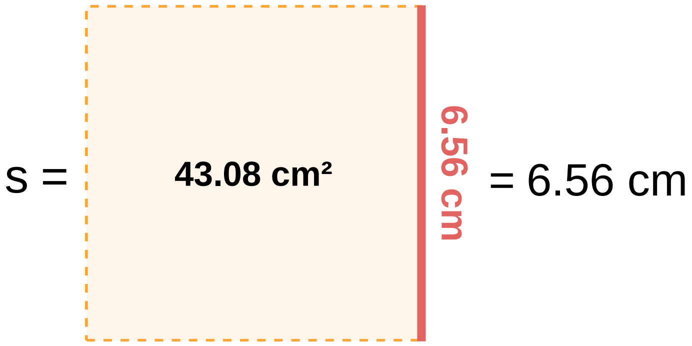


Figure 9.6: foo

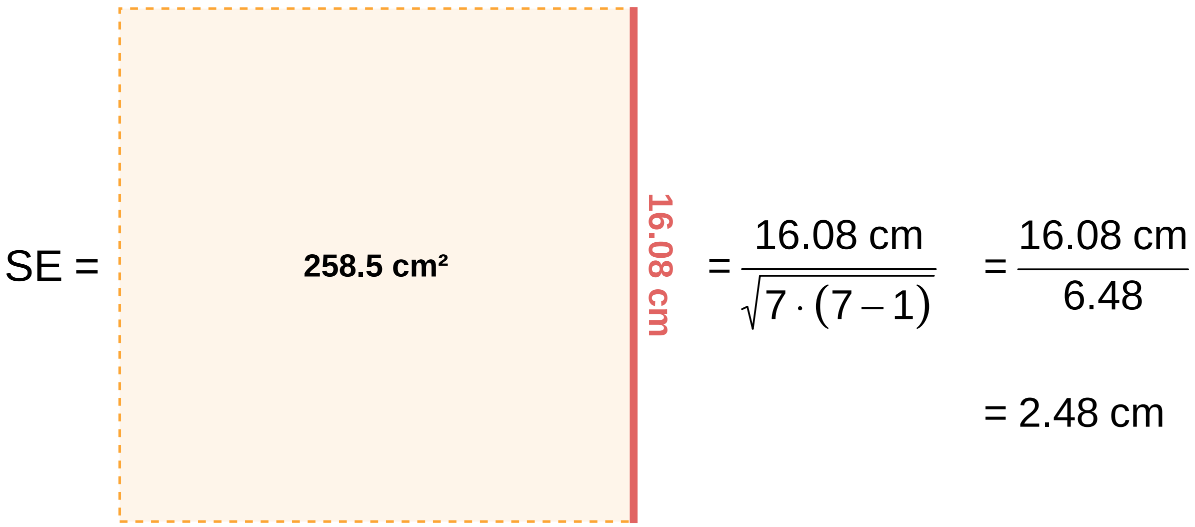


Figure 9.7: foo

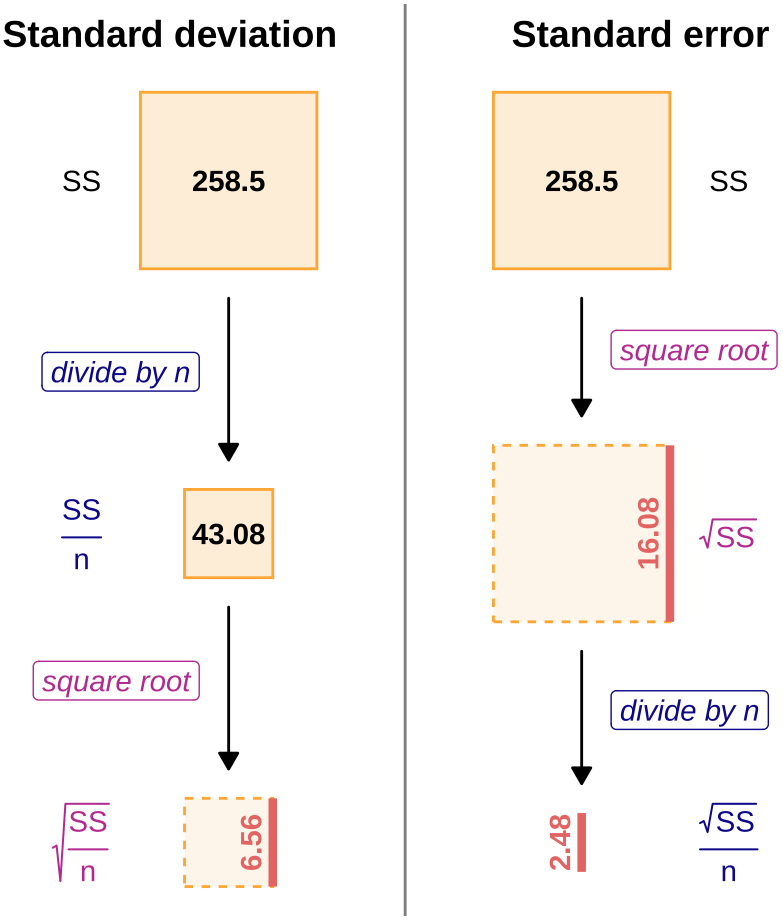


Figure 9.8: foo

9.5 Mean

9.6 Alternatives

Further tutorials and R packages on XXX

9.7 Glossary

term what does it mean.

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

- itemize with max. 5-6 words

9.9 Summary

References

10 Correlation and R^2

Last modified on 26. December 2025 at 19:12:18

> *“A quote.” — Dan Meyer*

10.1 General background

```r
r2_good_tbl <- tibble(weight = abs(rnorm(5, 3, 4)),
                      jumplength = 10 + 1.2 * weight + rnorm(5, 0, 3))

r2_good_fit <- lm(jumplength ~ weight, data = r2_good_tbl)

mean_good_cat_jump <- mean(r2_good_tbl$jumplength)

r2_good_plot_tbl <- r2_good_tbl |>
  mutate(sy = jumplength - mean(jumplength),
         e = residuals(r2_good_fit))

sum_good_sy <- (r2_good_plot_tbl$sy)^2 |> abs() |> sum() |> round(2)
sum_good_e <- (r2_good_plot_tbl$e)^2 |> abs() |> sum() |> round(2)
r2_good <- 1-(sum_good_e/sum_good_sy)


r2_bad_tbl <- tibble(weight = r2_good_tbl$weight,
                     jumplength = 9 + 0.5 * weight + rnorm(5, 0, 4))

r2_bad_fit <- lm(jumplength ~ weight, data = r2_bad_tbl)

mean_bad_cat_jump <- mean(r2_bad_tbl$jumplength)

r2_bad_plot_tbl <- r2_bad_tbl |>
  mutate(sy = jumplength - mean(jumplength),
         e = residuals(r2_bad_fit),
         e2 = e^2)
```

```
sum_bad_sy <- (r2_bad_plot_tbl$sy)^2 |> abs() |> sum() |> round(2)
sum_bad_e <- (r2_bad_plot_tbl$e)^2 |> abs() |> sum() |> round(2)
r2_bad <- 1-(sum_bad_e/sum_bad_sy)
```

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

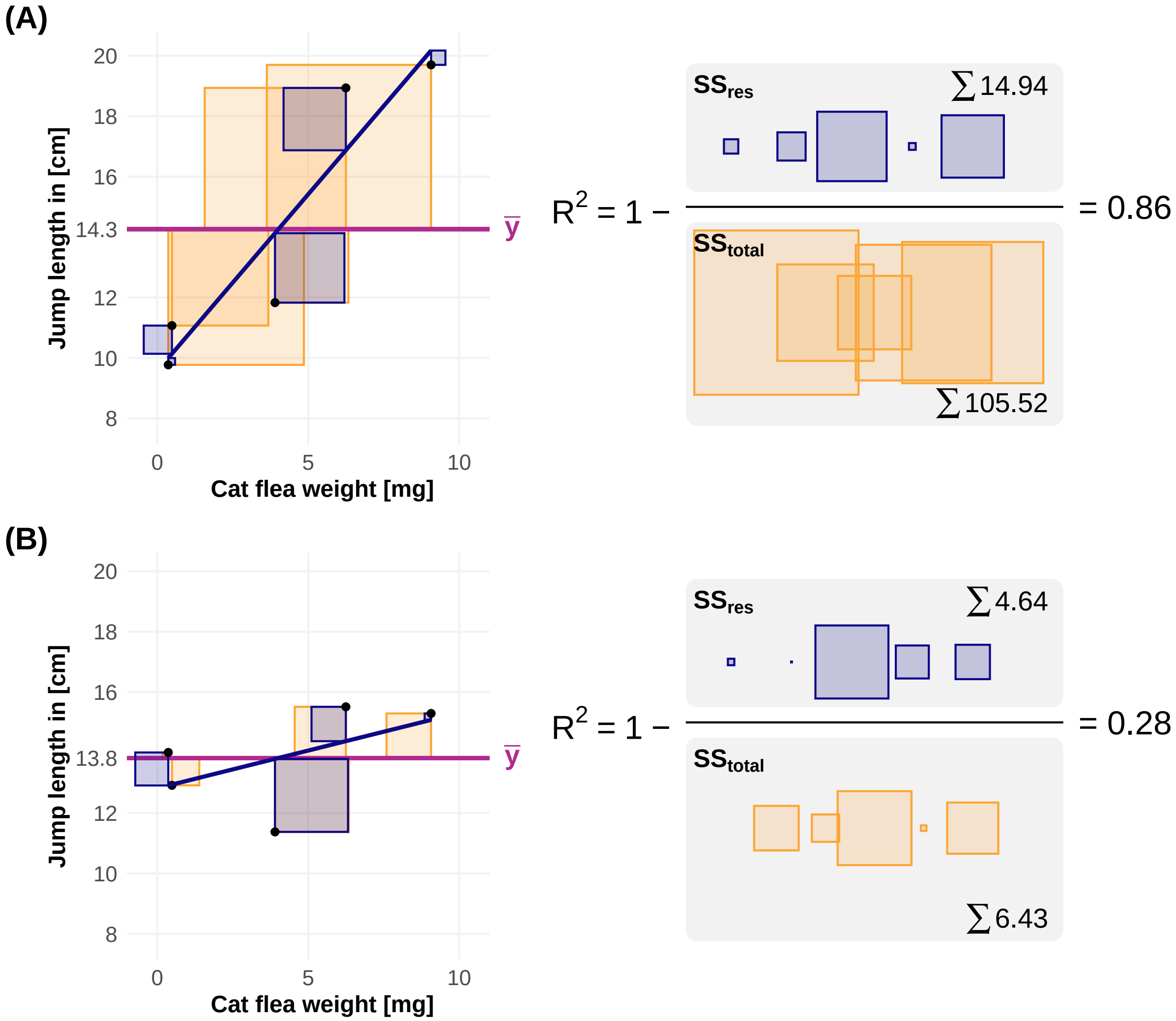


Figure 10.1: foo

```r
set.seed(20251226) #20251226
cor_high_tbl <- tibble(weight = abs(rnorm(5, 3, 4)),
                       jumplength = 10 + 1 * weight + rnorm(5, 0, 4)) |>
  mutate(sweight = weight - mean(weight),
         sjump = jumplength - mean(jumplength),
         ss_xy = sweight * sjump,
         ss_x = sweight^2,
         ss_y = sjump^2,
         sign_xy = ifelse(sign(ss_xy) == -1, "\U2012", "+"))
set.seed(202511)

sum(cor_high_tbl$ss_x)
```

```
[1] 33.95848
```

```r
sum(cor_high_tbl$ss_x) |> sqrt()
```

```
[1] 5.82739
```

```r
sum(cor_high_tbl$ss_y)
```

```
[1] 47.94158
```

```r
sum(cor_high_tbl$ss_y) |> sqrt()
```

```
[1] 6.923986
```

```r
sum(cor_high_tbl$ss_xy)
```

```
[1] 33.21142
```

```r
cor(cor_high_tbl$weight, cor_high_tbl$jumplength)
```

```
[1] 0.8231087
```

$$r = \frac{SS_{xy}}{\sqrt{SS_x} \cdot \sqrt{SS_y}}$$

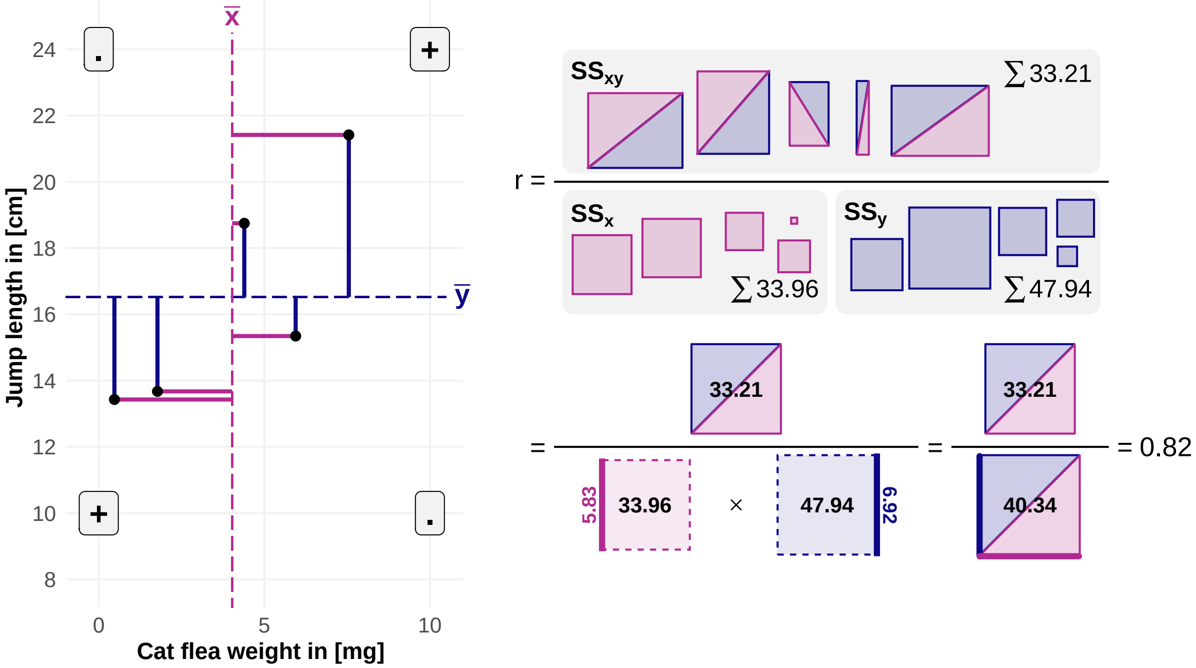


Figure 10.2: foo

10.2 Theoretical background

10.3 R packages used

10.4 Data

10.5 Alternatives

Further tutorials and R packages on XXX

10.6 Glossary

term what does it mean.

10.7 The meaning of "Models of Reality" in this chapter.

- itemize with max. 5-6 words

10.8 Summary

References

Part VI

Statistical modeling

Last modified on 18. December 2025 at 19:09:01

> *"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

11 Overview

Last modified on 18. December 2025 at 19:09:05

> *“A quote.” — Dan Meyer*

11.1 General background

11.2 Theoretical background

11.3 R packages used

11.4 Data

11.5 Alternatives

Further tutorials and R packages on XXX

11.6 Glossary

term what does it mean.

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

- itemize with max. 5-6 words

11.8 Summary

References

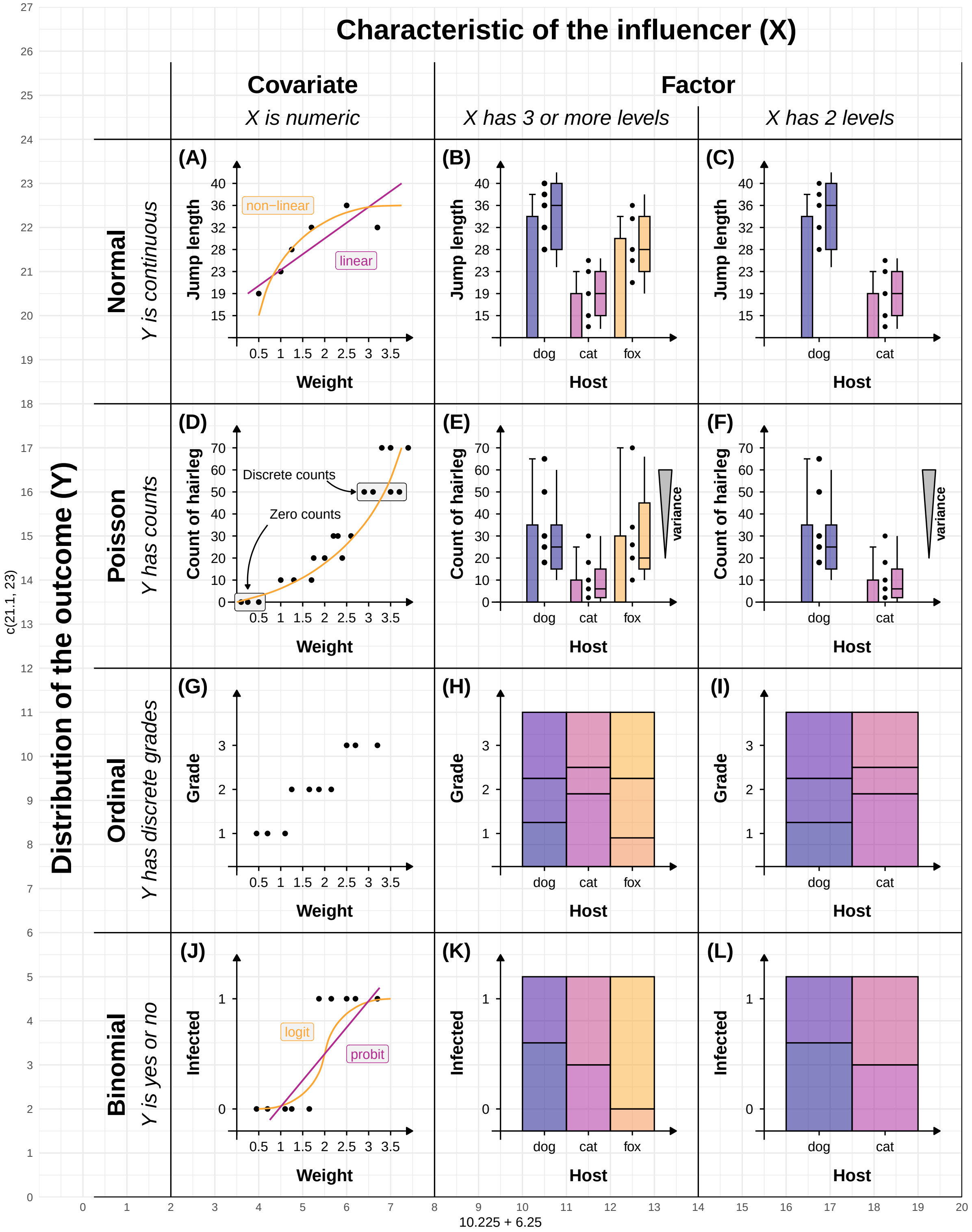


Figure 11.1: foo

Part VII

Template Preface

Last modified on 09. December 2025 at 10:27:51

> *"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

12 Template chapter

Last modified on 28. December 2025 at 17:34:59

> *“A quote.” — Dan Meyer*

12.1 General background

```r
set.seed(20251226) #20251226
cor_high_tbl <- tibble(weight = abs(rnorm(5, 3, 4)),
                       jumplength = 10 + 1 * weight + rnorm(5, 0, 4)) |>
  mutate(sweight = weight - mean(weight),
         sjump = jumplength - mean(jumplength),
         ss_xy = sweight * sjump,
         ss_x = sweight^2,
         ss_y = sjump^2,
         sign_xy = ifelse(sign(ss_xy) == -1, "\U2012", "+"))
set.seed(202511)

sum(cor_high_tbl$ss_x)
```

```
[1] 33.95848
```

```r
sum(cor_high_tbl$ss_x) |> sqrt()
```

```
[1] 5.82739
```

```r
sum(cor_high_tbl$ss_y)
```

```
[1] 47.94158
```

```r
sum(cor_high_tbl$ss_y) |> sqrt()
```

```
[1] 6.923986
```

```r
sum(cor_high_tbl$ss_xy)
```

```
[1] 33.21142
```

```r
cor(cor_high_tbl$weight, cor_high_tbl$jumplength)
```

```
[1] 0.8231087
```

```r
cov(cor_high_tbl$weight, cor_high_tbl$jumplength)
```

```
[1] 8.302855
```

```r
lm_fit <- lm(jumplength ~ weight, data = cor_high_tbl)
```

```r
vcov(lm_fit)
```

```
            (Intercept)     weight
(Intercept)   3.4921715 -0.6111907
weight       -0.6111907  0.1517616
```

12.2 Theoretical background

12.3 R packages used

12.4 Data

12.5 Alternatives

Further tutorials and R packages on XXX

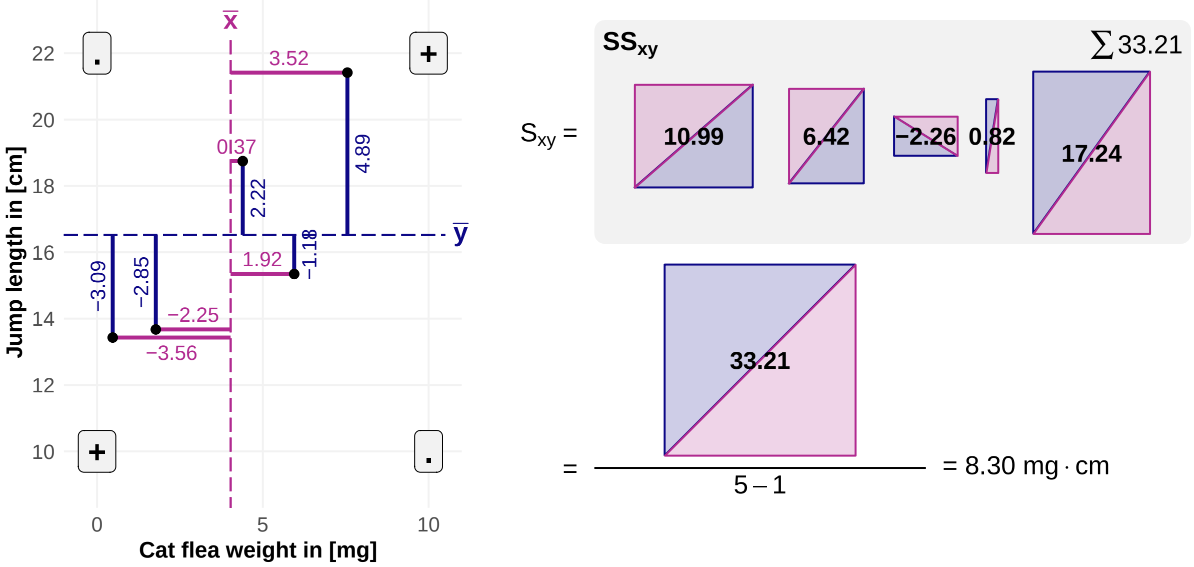


Figure 12.1: foo

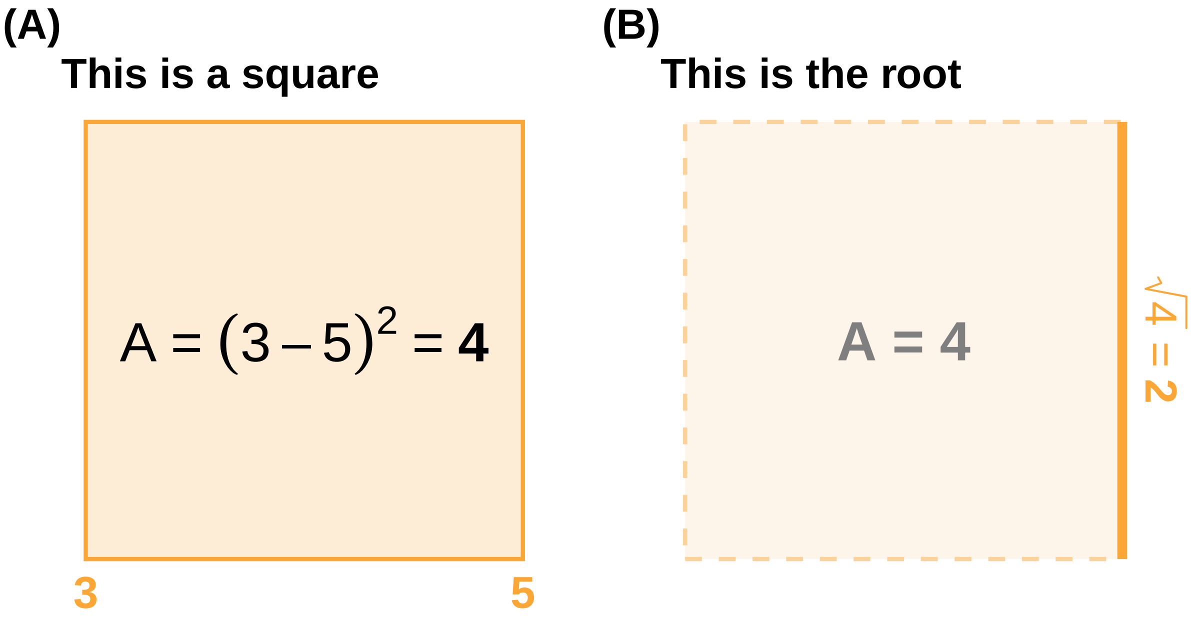


Figure 12.2: foo

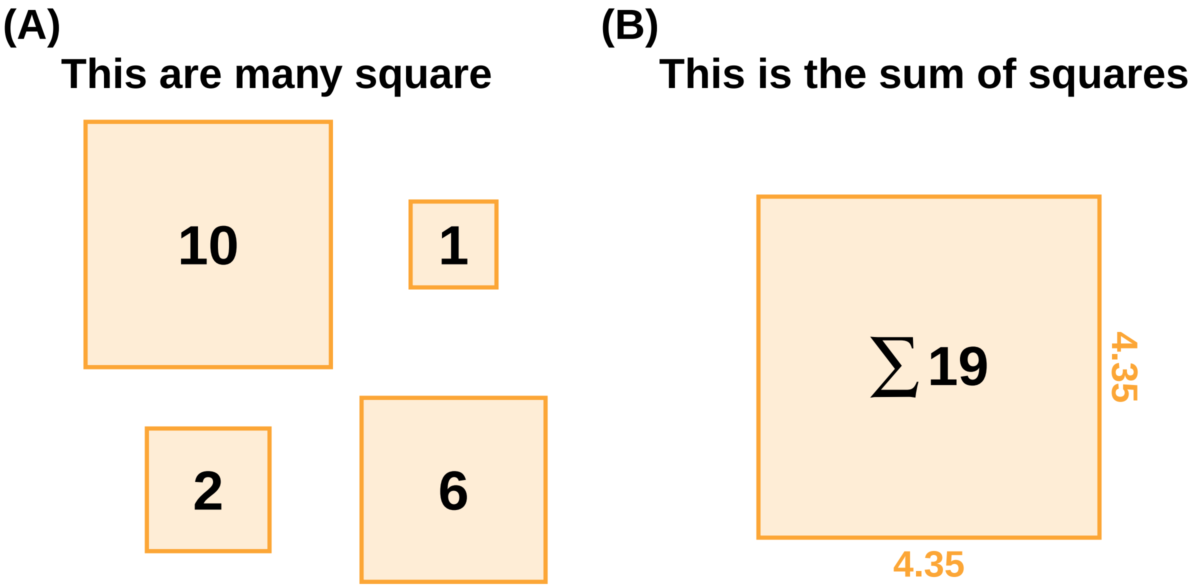


Figure 12.3: foo

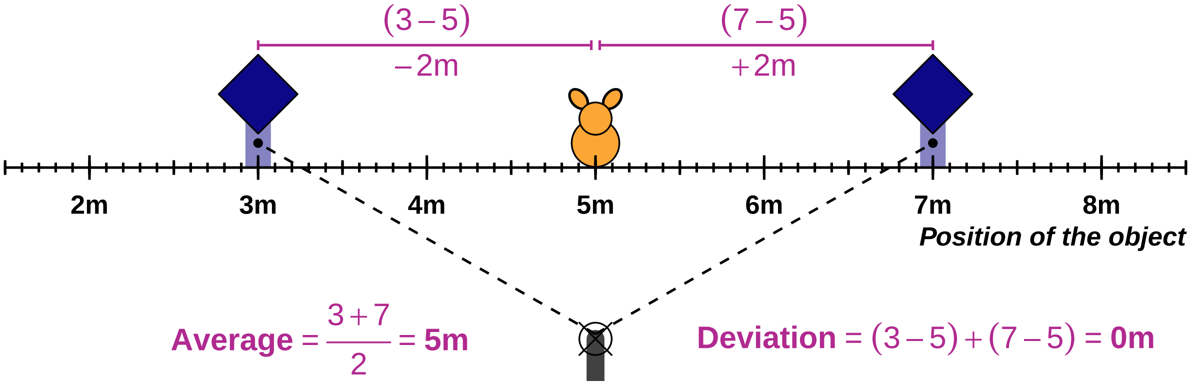

(3 – 5)
– 2m
(7 – 5)
+ 2m
2m
3m
4m
5m
6m
7m
8m
Position of the object
Average = $\frac{3+7}{2}$ = 5m
Deviation = (3 – 5) + (7 – 5) = 0m

Figure 12.4: foo

12.6 Glossary

term what does it mean.

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

- itemize with max. 5-6 words

12.8 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

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

A.2 Used theme of the visualitations

```r
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] Westfall PH, Arias AL. *Understanding Regression Analysis: A Conditional Distribution Approach*. Chapman; Hall/CRC; 2020.

[2] Smolin L. *Einstein's Unfinished Revolution: The Search for What Lies Beyond the Quantum*. Penguin; 2019.

[3] Feynman RP, Leighton R. *" Surely You're Joking, Mr. Feynman!": Adventures of a Curious Character*. Random House; 1992.

[4] Feynman RP. Cargo cult science. In: *The Art and Science of Analog Circuit Design*. Elsevier; 1998:55-61.

[5] 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.

[6] Everett III H. 'Relative state' formulation of quantum mechanics. *Reviews of modern physics*. 1957;29(3):454.

[7] Clarke AC. Clarke's third law on UFO's. *Science*. 1968;159(3812):255-255.

[8] Chalmers AF. *What Is This Thing Called Science?* Hackett Publishing; 2013.

[9] Feynman R. What is science. Published online 1966.

[10] Campbell NR. What is science? Published online 1952.

[11] Clark M, Berry S. *Models Demystified: A Practical Guide from Linear Regression to Deep Learning*. CRC Press; 2025.

[12] Cox D, Efron B. Statistical thinking for 21st century scientists. *Science advances*. 2017;3(6):e1700768.

[13] Deutsch D. *The Beginning of Infinity: Explanations That Transform the World*. penguin uK; 2011.

[14] McCullagh P. What is a statistical model? *The Annals of Statistics.* 2002;30(5):1225-1310.

[15] Appleton DR. What do we mean by a statistical model? *Statistics in medicine.* 1995;14(2):185-197.

[16] Hand D. What is the purpose of statistical modeling? *Harvard Data Science Review.* 2019;1(1).

[17] Spanos A. Where do statistical models come from? Revisiting the problem of specification. *Lecture Notes-Monograph Series.* Published online 2006:98-119.

[18] Gilchrist W. *Statistical Modelling.* Wiley Chichester; 1984.

[19] Wickham H, Çetinkaya-Rundel M, Grolemund G. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* " O'Reilly Media, Inc."; 2023.

[20] Wickham H, Henry L. *Purrr: Functional Programming Tools.*; 2025. https://purrr.tidyverse.org/

[21] Dormann CF, Ellison AM. *Statistics by Simulation: A Synthetic Data Approach.* Princeton University Press; 2025.

[22] Wickham H. *Conflicted: An Alternative Conflict Resolution Strategy.*; 2023. https://conflicted.r-lib.org/

[23] Rinker T, Kurkiewicz D. *Pacman: Package Management Tool.*; 2012. https://github.com/trinker/pacman