Rethinking Companion

Wade VanderWright

2022-05-18

Contents

Chapter 1

The Golem of Prague

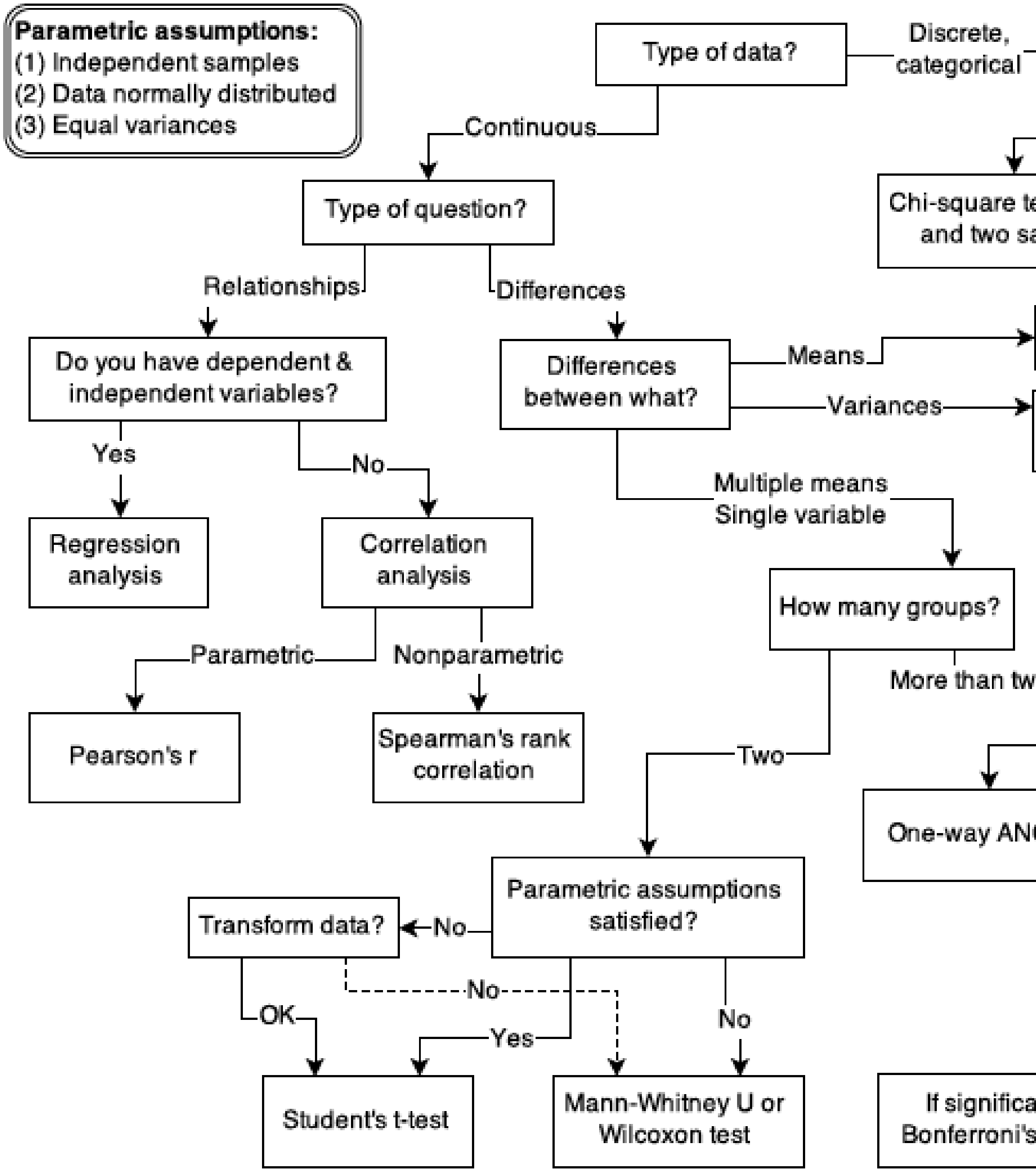
This is a *companion* book written in Markdown for McElreath's *Statistical Rethinking* (2020). You can set up your R console by running:

```
install.packages(c("coda","mvtnorm","devtools","dagitty"))
library(devtools)
devtools::install_github("rmcelreath/rethinking")
```

1.1 Statistical golems

The Golem of Prague and statistical golems (models) are powerful but lack wisdom. As McElreath tells us, there are many kinds of golems and figuring out how to build the one you need to carry out the task at hand can be tricky.

Figure 1.1


Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances
Type of data?
Discrete,
categorical
Continuous
Chi-square te
and two sa
Type of question?
Relationships
Differences
Do you have dependent &
independent variables?
Differences
between what?
Means
Variances
Yes
No
Multiple means
Single variable
Regression
analysis
Correlation
analysis
How many groups?
Parametric
Nonparametric
More than tw
Pearson's r
Spearman's rank
correlation
Two
One-way AN
Parametric assumptions
satisfied?
Transform data?
No
No
OK
Yes
No
Student's t-test
Mann-Whitney U or
Wilcoxon test
If significa
Bonferroni's

In addition, novel research often requires novel methods and the researchers may have to stray from the common tests to engineer their own golems.

1.2 Statistical Rethinking

> A lot can go wrong with statistical inference, and this is one reason that beginners are so anxious about it. When the goal is to choose a pre-made test from a flowchart, then the anxiety can mount as one worries about choosing the "correct" test.

More work is needed to ensure researchers understand all the moving parts of their golems and how to interpret their results.

1.2.1 What are we trying to do with the golems?

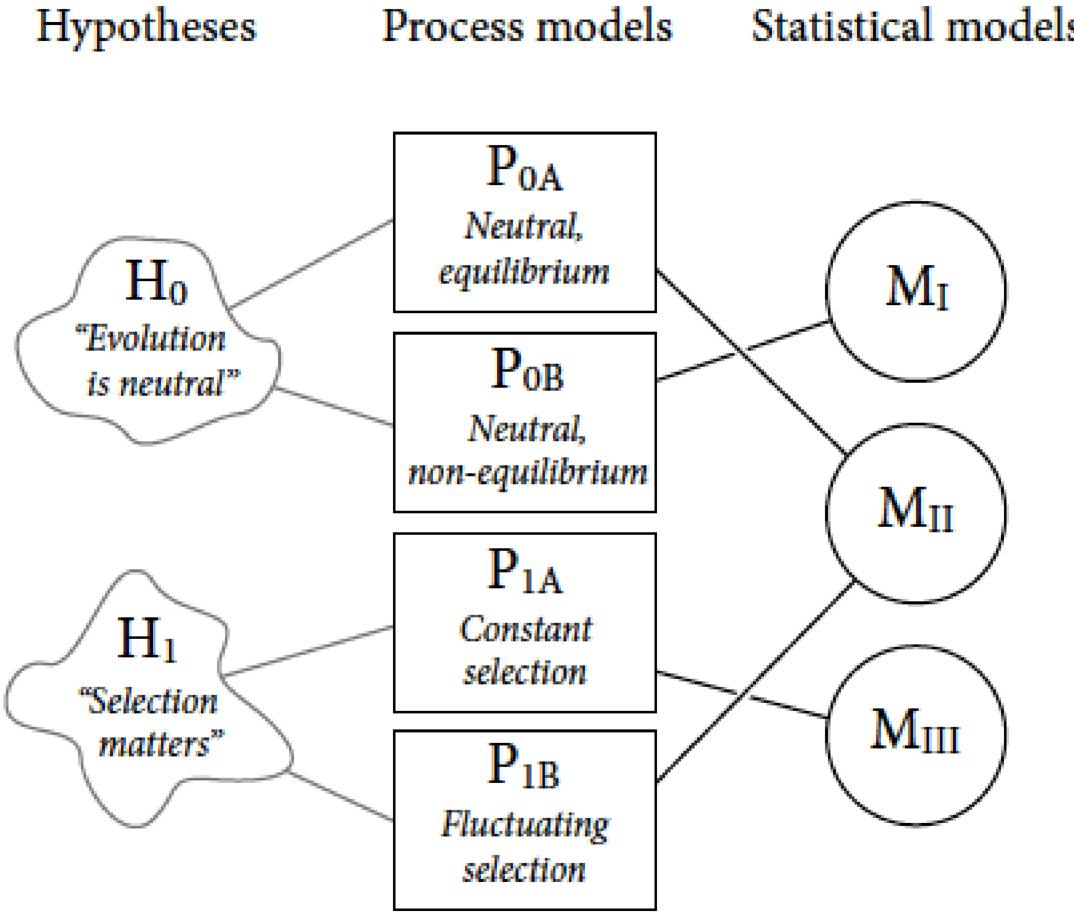
The popular belief is that we need to create models that use statistical means to test the null hypothesis.

Two reasons why deductive falsification doesn't work:

> 1. Hypotheses are not models. The relations among hypothese and different kinds of models are complex. Many models correspond to the same hypothesis, and many hypotheses corresponf to asingle model. This makes strict falsification impossible.

All models are false, but some are useful.

Figure 1.2



Two opposing hypothesis for evolutionary change:

H_0: Neutral theory (random mutation and drift)

H_1: Natural selection (fitness leads to observed change)

Process models for each hypothesis:

P_{0a}: steady state in time (null)

P_{0b}: fluctuations in population size through time

P_{1a}: selection favours the same alleles through time

P_{1b}: selection preference fluctuates through time (different alleles)

Statistical Models:

M_i: unique to P_{0b}

M_{ii}: Power law in the data (frequency) shared expectation of P_{0a} *and* P_{1b}

M_{iii}: unique to P_{1a}

Note that all process models contain time, solidifying directionality

2. Measurements matter. Even when we think the data falsify a model, another observer will debate our methods and measures. They don't trust the data. Sometime sthey are right.

The colour of swans

Before Australia was discovered, all swans were white and no number of observations could prove this fact to be true.

H_0: All swans are white

Australia had black swans, which instantly makes H_0 false.

Remember, observations are prone to error and hypotheses are quantitative rather than discrete.

> "At the edges of scientific knowledge, the ability to measure a hypothetical phenomenon is often in question as much as the phenomenon itself."

1.3 Tools for golem engineering

> You'll wreck Prague eventually, you just need to notice the destruction.

We want our models to be able to design inquiry, extract information from data, and make predictions. To do this we will need:

1. Bayesian data analysis
2. Model comparison
3. Multilevel Models
4. Graphical causal models

1.3.1 Bayesian data analysis

Bayesian data analysis takes questions in the form of a model and produces logical probability distributions of the answer. This represents plausibility.

1.3.2 Model comparison and predictions

Model comparison is often thought of in terms of ‘which model will make the best predictions?’ Two tools for this are Cross-validation and Information Criteria.

Complex models usually make worse predictions than simple ones due to *overfitting*. The smarter the golem, the dumber its predictions. Fitting is easy; prediction is hard.

1.3.3 Multilevel models

1.3.4 Graphical causal models

1.4 Summary

Session Info

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_CA.UTF-8/en_CA.UTF-8/en_CA.UTF-8/C/en_CA.UTF-8/en_CA.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.1.1  magrittr_2.0.1  fastmap_1.1.0   bookdown_0.26
##  [5] htmltools_0.5.2 tools_4.1.1     rstudioapi_0.13 yaml_2.2.1
##  [9] stringi_1.7.3   rmarkdown_2.14  knitr_1.33      stringr_1.4.0
## [13] digest_0.6.27   xfun_0.30       rlang_0.4.11    png_0.1-7
## [17] evaluate_0.14
```

Chapter 2

Small worlds and large worlds

Every model has two parts: small world and large world. *The small world* is within the model itself and *the large world* is the broader world we want the model to be applied to. In the small world, everything is defined and there isn't much room for pure surprises. The large world has more opportunities for unforeseen events and because the small world is an incomplete representation of the large world, mistakes are expected. The goal is to create small worlds (models) that approximate reality so that they perform well in the large world.

2.1 The garden of forking data

Bayesian inference is the counting and comparing of possibilities. At each point where a 'decision' may split the path, bayesian inference evaluates each path and eliminates the paths that are not congruent with the data fed into the model.

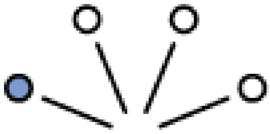
2.1.1 Counting possibilities

Marble Example

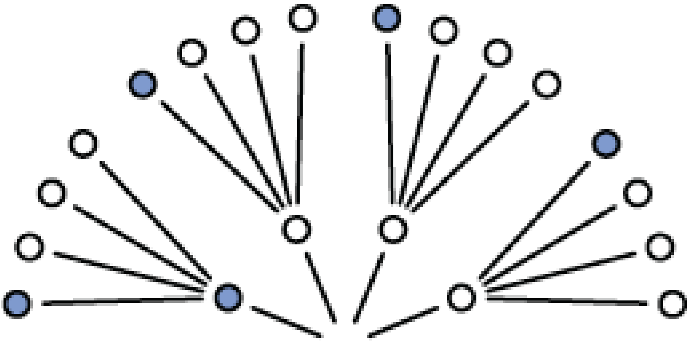
There is a bag of four marbles of two colours (blue and white). This means that there could be 5 possibilities (conjectures); 4:0 white, 3:1 white, 2:2 split, 3:1 blue, and 4:0 blue.

A sequence of three marbles is pulled from the bag, one at a time, and returned to the bag (repeated sampling)
We get blue, white, blue.

Considering a 3:1 white scenario, on the first draw you could get a blue marble or three white marble draws



Expanding out one more draw (layer) we can expect the same possibilities because the first marble is replaced before the second draw



Expanding one more time gives us the final garden of 64 possibilities (4^3; 4 marbles with 3 draws)

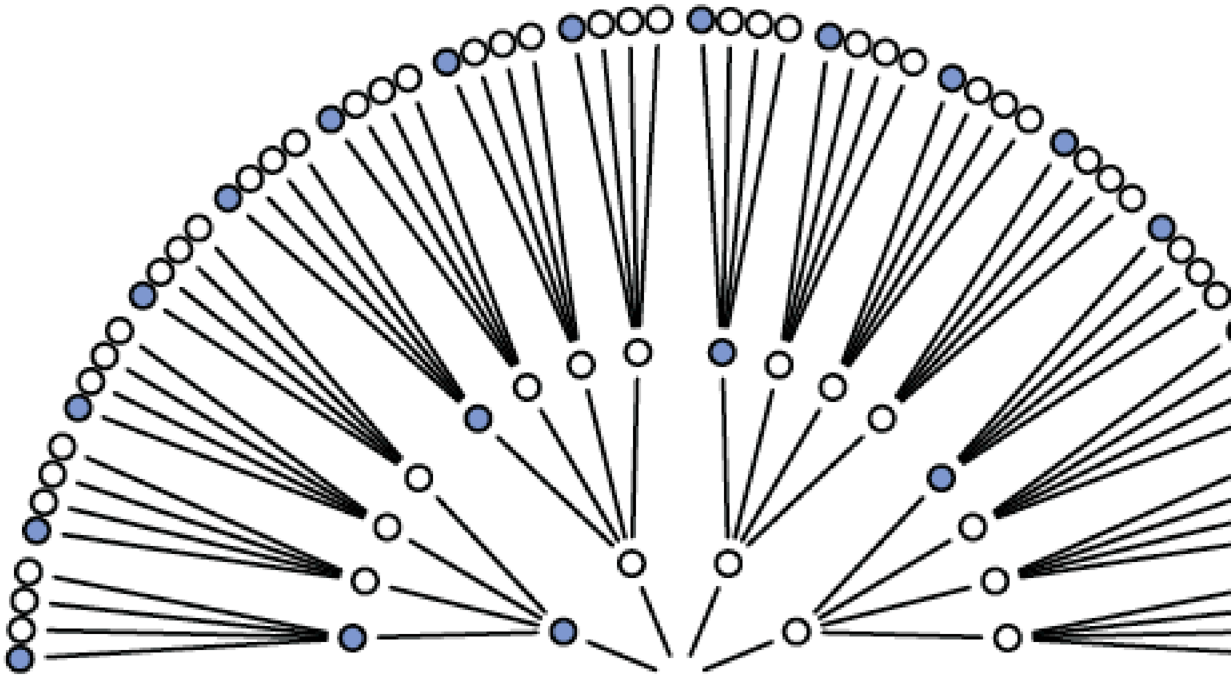


Figure 2.2. The 64 possible paths generated by assuming the bag c
one blue and three white marbles.

Now recall our draws were blue, white, blue so we can trim the paths that are not congruent with the draws

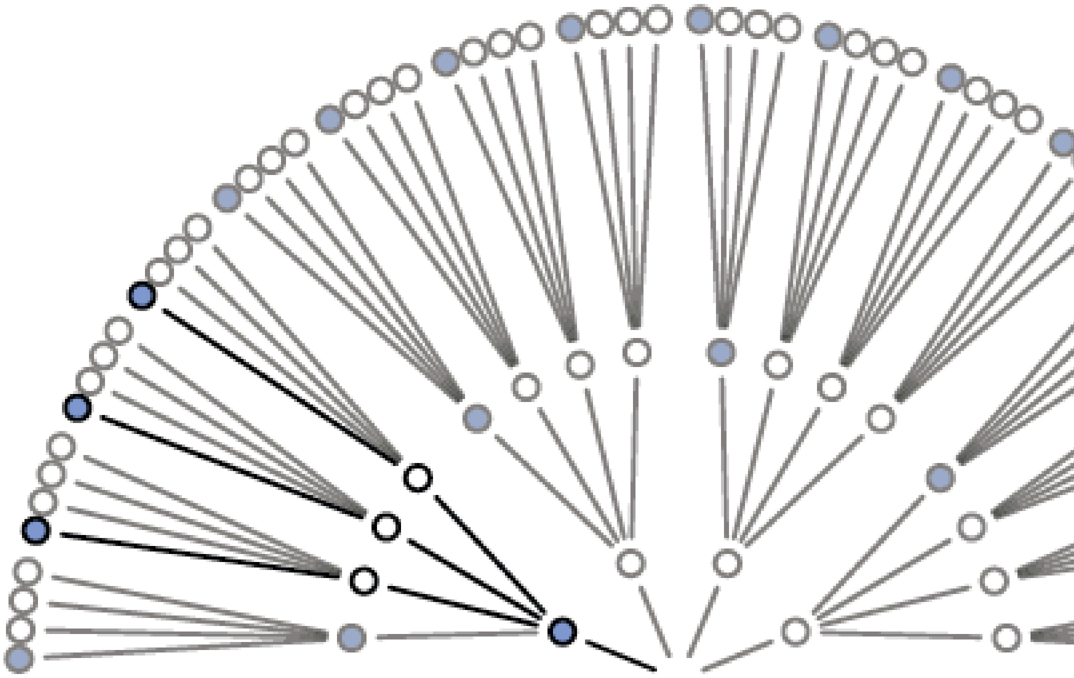
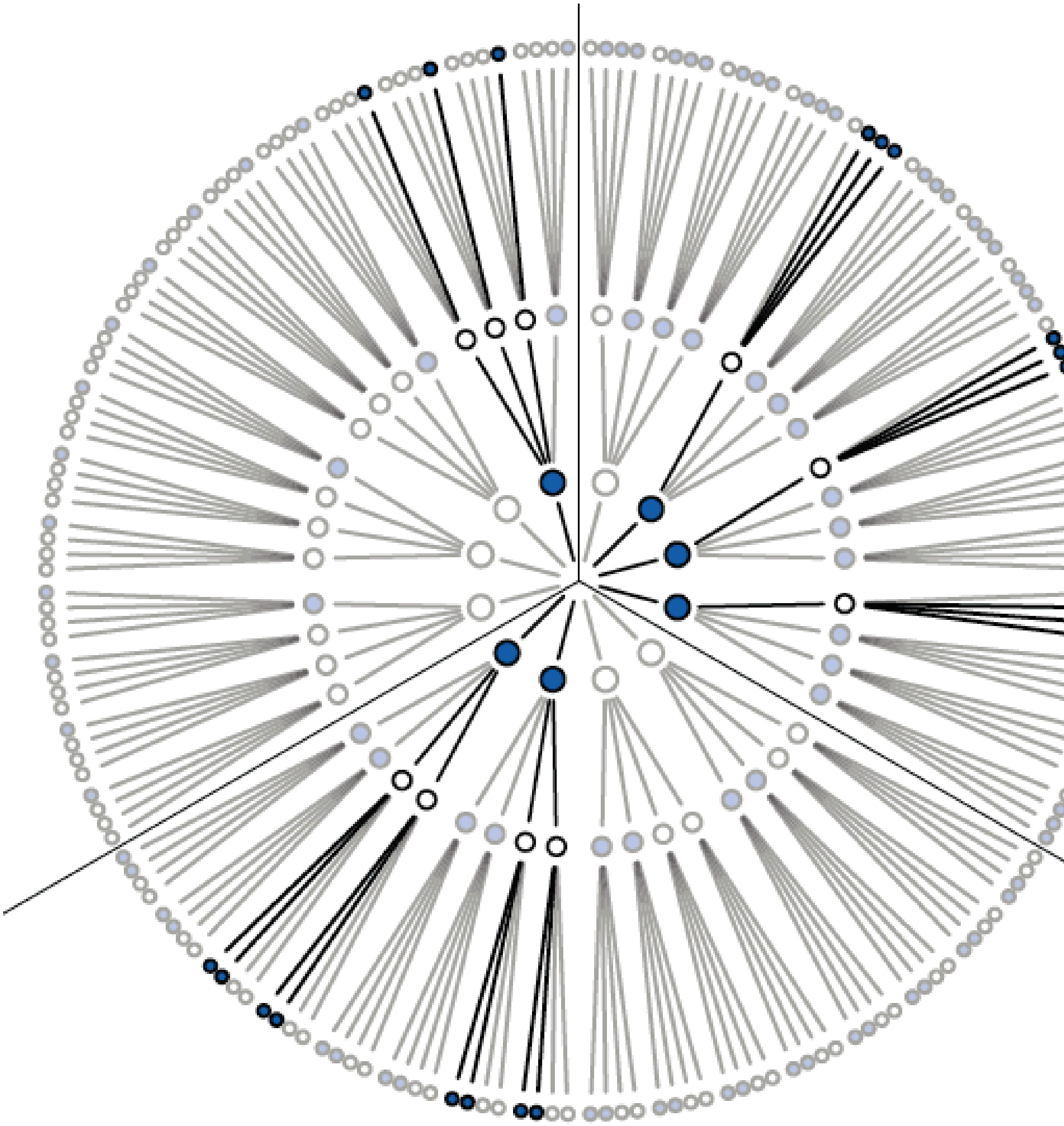


Figure 2.3. After eliminating paths inconsistent with t
quence, only 3 of the 64 paths remain.

We can also trim other possibilities like all white marbles or all blue marbles because we drew both colours from the bag. Putting our 3:1 white, 2:2 split, and 3:1 blue possibilities together would look something like this



You can see that there are different numbers of unique paths to get our observed result
3:1 white has 3 paths
2:2 split has 8 paths
3:1 blue has 9 paths

We will call these counts our priors.

2.1.2 Combining other information

Suppose we make another marble draw and it is blue. We then count the ways each of our marble possibilities could create this new result

3:1 white has 1 paths
2:2 split has 2 paths
3:1 blue has 3 paths

Mutiplying by the prior counts gives us:

3:1 white has (3x1) 3 paths
2:2 split has (8x2) 16 paths
3:1 blue has (9x3) 27 paths

and suggests that our 3:1 blue possibility is more plausible with the new information.
Note that prior data and new data don't have to be of the same type

Chapter 3

Sampling the imaginary

Chapter 4

Geocentric models

Chapter 5

The many variables & the spurious waffles

Chapter 6

The haunted dag & the casual terror

6.1 Equations

Here is an equation.

$$f\left(k\right)=\binom{n}{k}p^k\left(1-p\right)^{n-k} \tag{6.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (6.1).

6.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 6.1.

Theorem 6.1. *For a right triangle, if c denotes the* length *of the hypotenuse and a and b denote the lengths of the* ***other*** *two sides, we have*

$$a^2+b^2=c^2$$

Read more here https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html.

6.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html

Chapter 7

Ulysses' compass

7.1 Publishing

HTML books can be published online, see: https://bookdown.org/yihui/bookdown/publishing.html

7.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

7.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `gitbook` uses the same social sharing data across all chapters in your book-all links shared will look the same.

Specify your book's source repository on GitHub using the `edit` key under the configuration options in the `_output.yml` file, which allows users to suggest an edit by linking to a chapter's source file.

Read more about the features of this output format here:

https://pkgs.rstudio.com/bookdown/reference/gitbook.html

Or use:

```
?bookdown::gitbook
```

Chapter 8

Conditional manatees

Chapter 9

Markov chain monte carlo

Chapter 10

Big entropy and the generalized linear model

Chapter 11

God spiked the integers

Chapter 12

Monsters and mixtures

Chapter 13

Models with memory

Chapter 14

Adventures in covariance

Chapter 15

Missing data and other opportunities

Chapter 16

Generalized linear madness

Chapter 17

Horoscopes