Statistical Intuition for Modern Biologists

Isaac Vock

2024-05-07

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

Part I

Necessary Introduction

1 Introduction to RNA-seq

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 Introduction to R

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

3 Introduction to Probability

Why collect replicates of the same exact data? The (in)famous definition of insanity, "Doing something over and over again and expecting a different result", would seem to classify replication as crazy. Except, anyone who has ever collected any kind of data has an intuition for why replication is important: Each replicate is always different from the last. Somehow, despite following a precise protocol in a particular lab with the same reagents, pipettes, measuring devices, etc., each replicate yields a different result. **There is randomness in your data**.

This randomness throws a wrench in naive data analyses. In biology, we are often measuring something (e.g., the concentration of an interesting molecule, the viability of an organism, etc.) in two or more "conditions" (e.g., with and without drug) and assessing whether or not any of our measurements differ from one another. If every replicate of an experiment yielded the same data, this task would be trivial. Just take one measurement in each condition, compare the measurements, and draw conclusions. If the measurements differ, the thing you are measuring differs. If the measurements are identical, the thing you are measuring is unaffected by your perturbations. The randomness of data forces us to consider an alternative conclusion. What if our measurements differ, not because the thing we are measuring is changing, but because our measurements fluctuate from experiment-to-experiment? How do we know what differences are real and what differences are the result of this randomness?

The answer discussed in this chapter is **probability theory**. Probability theory is a field of mathematics that gives us the tools to think about and quantify this randomness. This chapter will introduce you to the key instrument provided by this field: the probability distribution. With this tool, we will be able to describe and better understand the random fluctuations that plague our data.

3.1 The Many Flavors of Randomness

Trajectory: 1) In biology, everything is complicated and every system you study seems unique. 2) To tackle this complexity, we group things into bins based on important characteristics that they share, and then try and understand the patterns that govern each bin. 3) Similarly in statistics, every kind of data seems different and to suffer from its own sources of variance. 4) We notice though, that there are patterns in the types of randomness we commonly see. We thus focus on understanding these common patterns and apply them to understanding our unique data. 5) The patterns in randomness that we observe are coined "probability

distributions". Think of them as functions, which take as input a number, and provide as output the probability of observing that number (or a tiny range of numbers around that number). 6) These functions could take any shape you dream up, but as mentioned, particular patterns crop up all of the time. In this class, we are going to explore these patterns and understand what gives rise to them. 7) One way to think about a probability distribution is as specifying a strategy to randomly generate data that follows a particular pattern of randomness. The height of the function at any point determines how likely you should be to draw that particular number. 8) Such random number generators exist in R, and generate data that follow a number of the most common patterns we see in the real world. This exercise walks you through each of these and helps you make connections between them, and to begin to understand the origin of the patterns you'll see.

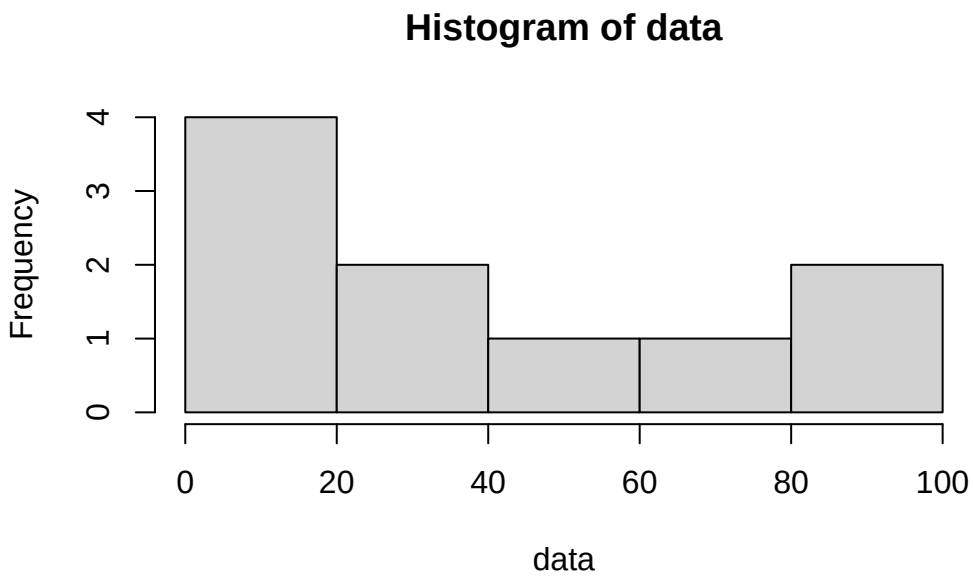
Random number generators to assess: - rnorm() - rbinom() - runif() - rbeta() - rgamma() - rexp() - rlnorm() - rnbinom() - rgeom() - rhypergeom() - rpois()

3.2 Appendix: Probability Distributions

When I claim that "there is randomness in your data", what does that mean? For some people, the term "randomness" implies complete unpredictability. Such people would interpret my claim to mean that every time you collect a new replicate, any value for the thing you are measuring is fair game and equally likely. "You got 100 reads from the MYC gene in your last RNA-seq dataset? Well don't be surprised if you get 1000, or 10000, or 0 reads next time!" You may call this **uniform randomness**. You can generate such data right here in R, using the `runif()` function:

```
data <- runif(n = 10, min = 0, max = 100)

hist(data)
```



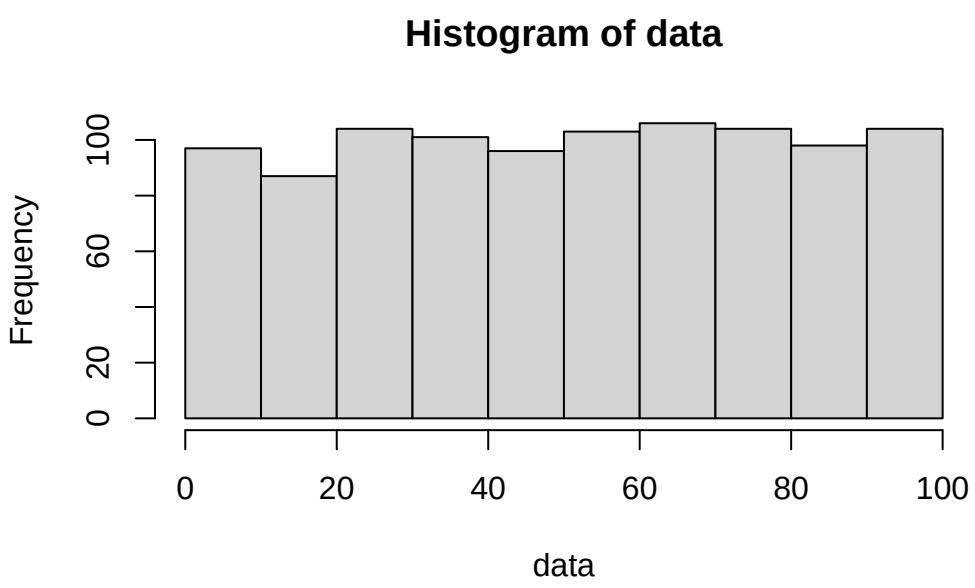

```
# Check out the individual data points
print(data)
```

```
 [1] 18.91652 43.27310 16.09708 88.82375 10.85821 31.91279 16.11527 33.51403
 [9] 84.15258 67.34352
```

`runif()` has three parameters: 1) `n` specifies the number of numbers to generate, 2) `min` specifies the minimum number it could possibly generate, and 3) `max` specifies the maximum number it could possibly generate. In the above example, I am thus creating 10 numbers between 0 and 100, with every number in between being equally likely to pop out. Generate a lot more numbers and this uniform pattern of appearance becomes much more clear:

```
data <- runif(n = 1000, min = 0, max = 100)

hist(data)
```



While this definition of randomness is intuitive, it can't be the only type of randomness. If RNA-seq data were this random, it would be useless! There is nothing to learn from measurements that can take on any value with equal probability.

Thus, to describe all of the kinds of randomness we see in the real world, it is important to expand our definition beyond uniform randomness. Enter the **probability distribution**. A probability distribution is like a function in R. It takes as input a number (or maybe a set of numbers), and provides as output, the probability of seeing that number. For uniformly random data this might look something like:

uniform_distribution <- function(data, min = 0, max = 100){

if(data >= min & data <= max){

```
output <- 1
```

}else{

```
output <- 0
```

}

return(output)

} "'it ad

That is to say, as long as the data is within the bounds of what is possible, it has the same probability of occuring; you get the same number out from this function. This function would make mathematicians

4 Introduction to Statistical Modeling

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

Part II

Popular Methods

5 Hypothesis Testing

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

6 Linear Modeling

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

7 Dimensionality Reduction

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

8 Clustering and Mixture Modeling

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

Part III

Statistics in the Wild

9 Practical Bioinformatics

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

10 Analysis of an RNA-seq dataset

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Knuth, Donald E. 1984. "Literate Programming." *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.