

The (many) benefits of simulated data

Ben Lambert¹

`ben.c.lambert@gmail.com`

¹University of Oxford

Wednesday 23rd June, 2021

Course plan

- 9.30am-10.15am: lecture, “The benefits of simulated data”
- 10.45am-1pm: practical
- 2pm-2.30pm: lecture, “Reproducible data analysis”
- 2.45pm-5pm: practical

Lecture content

- How to create useful methods (using simulation)
- Two key problems with statistical significance testing (explored via simulation in problem sets)
- Simulated data for inference
- Simulated data for experimental design

Outline

- 1 How to create useful methods
- 2 Two key problems with statistical significance testing
- 3 The use of simulated data in inference

- 1 How to create useful methods
- 2 Two key problems with statistical significance testing
- 3 The use of simulated data in inference

Null hypothesis testing

Assume we have a null hypothesis H_0 for how data are generated. For example, suppose:

$$X_i \sim \text{normal}(\theta, 1), \quad (1)$$

where $H_0 : \theta = 0$ versus an alternative hypothesis (say)
 $H_1 : \theta < 0$.

p -values

In statistical hypothesis testing, the p -value is the probability of observing something as least as extreme as the observed test statistic, $T(X)$:

$$p = \mathbb{P}(T(X^{\text{rep}}) \leq T(X)), \quad (2)$$

where if $H_0 : \theta = 0$,

$$X_i \sim \text{normal}(0, 1). \quad (3)$$

and we could have $X^{\text{rep}} = (X_1, X_2, \dots, X_N)$ and

$$T(X^{\text{rep}}) = \frac{1}{N} \sum_{i=1}^N X_i. \quad (4)$$

Statistical test size

- Reject H_0 : $p \leq \alpha$,
- Do not reject H_0 : $p > \alpha$.

Here,

$$\alpha = \mathbb{P}(\text{conclude } H_0 \text{ is false} | H_0 \text{ is true}) \quad (5)$$

is known as the size of a statistical test.

Statistical power

Suppose some alternative hypothesis $H_1 : \theta = \theta_1$ is true, then:

$$\text{power} = \mathbb{P}(\text{reject } H_0 | H_1 \text{ is true}) \quad (6)$$

So power relates to a **specific** alternative hypothesis, H_1 ; a test that's good for one H_1 may not be good at many others.

It's also typically defined relative to a given α value: for example, “the power to reject the null against specific H_1 using a statistical significance of α ”.

Designing methods

- Many of you will create methods for use by others;
- Important to ensure this is done responsibly so it can be replicated: good software testing and comprehensively documented;
- As important is to ensure that the methods are **useful**.

How to create and publish useful tests

Whilst there are many types of method, here we use statistical tests as a case study in ensuring usefulness.

A statistical test is useful if:

- 1 Its α behaves as it should under the distribution(s) defined by the null hypothesis;
- 2 It is powerful across a range of likely to be encountered H_1 s;
- 3 You have determined and communicated the H_1 s for which it doesn't work.

\implies can use simulation to handle all the above!

1. Checking α

I wrote the following imprecise statement for the null distribution of a single data point:

$$X_i \sim \text{normal}(0, 1). \quad (7)$$

There are a number of ways this could be true. For example,

$$X_i \overset{i.i.d.}{\sim} \text{normal}(0, 1). \quad (8)$$

Or (say),

$$X_i = \rho X_{i-1} + \epsilon_i, \quad (9)$$

where $|\rho| < 1$ and $\epsilon_i \overset{i.i.d.}{\sim} \text{normal}(0, \sqrt{1 - \rho^2})$.

Question: does your α behave as expected under these ranges? Or do you need to be more specific when defining H_0 .

2. and 3. checking power under a variety of H_1 s

Assume null distribution: $X \stackrel{i.i.d.}{\sim} \text{normal}(0, 1)$. There are a variety of alternative hypotheses:

- $H_1 : X \sim \text{normal}(-1, 1)$
- $H_1 : X \sim \text{normal}(0, 1.5)$
- $H_1 : X \stackrel{\text{non } i.i.d.}{\sim} \text{normal}(0, 1)$
- $H_1 : X \sim \text{Student-t}(\dots)$
- $H_1 : X \sim \text{skew-normal}(\dots)$
- $H_1 : X \sim \text{multimodal-normal}(\dots)$

Note, if all the above were relevant, you should communicate power results across all of these: good and bad.

- 1 How to create useful methods
- 2 Two key problems with statistical significance testing
- 3 The use of simulated data in inference

Statistical significance is not practical significance

Suppose two treatments aimed at increasing personal income¹:

- Treatment 1: estimated to increase annual earnings by \$10 with a standard error of \$2.
- Treatment 2: estimated to increase annual earnings by \$10,000 with a standard error of \$10,000.

Only treatment 2 has the potential to impact the real world but is not statistically significant.

⇒ make decisions on practical utility based on changes to predictive power.

¹From Gelman, Hill, Vehtari, 2021, *Regression and Other Stories*.

Statistical significance testing naturally leads to overestimation

For an estimate, $\hat{\theta}$, to be statistically significant, it must pass some threshold:

- Threshold higher for lower power tests;
- Threshold increases with the noisiness of the data.

Therefore the weaker the test and the noisier the data,

$$\mathbb{P}(\hat{\theta} > \theta | p < 0.05) \tag{10}$$

is higher (and can be really high: see problem set).

- 1 How to create useful methods
- 2 Two key problems with statistical significance testing
- 3 The use of simulated data in inference

Example model: Lotka-Volterra

Describe population dynamics of prey $x(t)$ and predator $y(t)$:

$$\frac{dx}{dt} = \alpha x - \beta xy \quad (11)$$

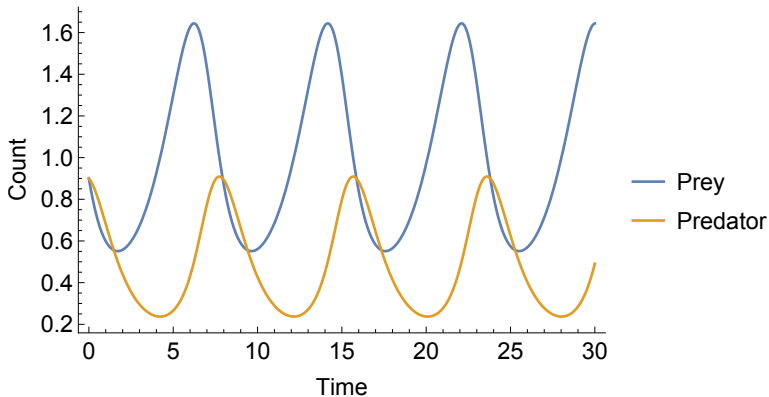
$$\frac{dy}{dt} = \delta xy - \gamma y \quad (12)$$

with $x(0) = x_0$ and $y(0) = y_0$.

Oscillatory dynamics

Assuming:

$$\alpha = 2/3, \beta = 4/3, \gamma = 1, \delta = 1, x(0) = 0.9, y(0) = 0.9 \quad (13)$$



Inference problem

Problem: Given prey series: $(x(0), x(10), x(20), x(30))$,
can we infer (β, γ) ?

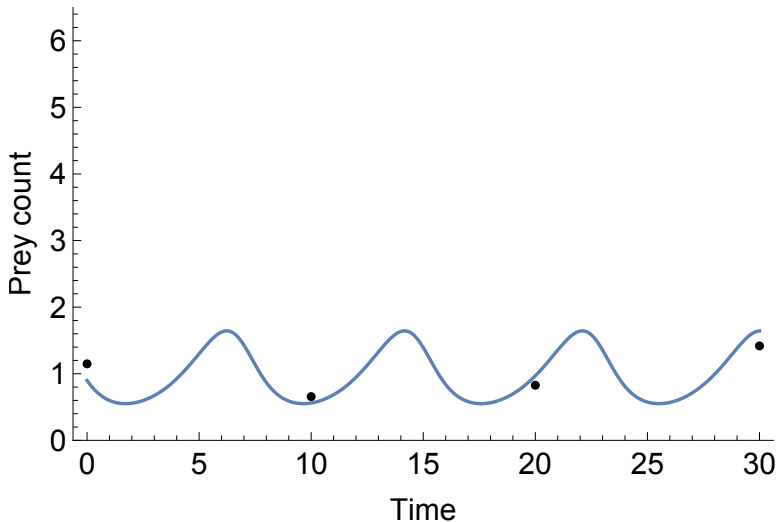
Answer: try inference for simulated data! Here, we assume
same set of parameters as before and

$$\tilde{x}(t) \stackrel{i.i.d.}{\sim} \text{normal}(x(t), 0.3), \quad (14)$$

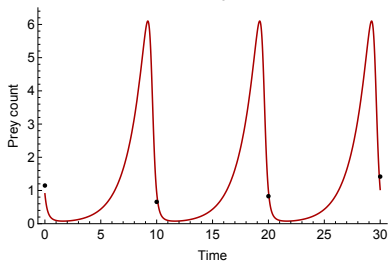
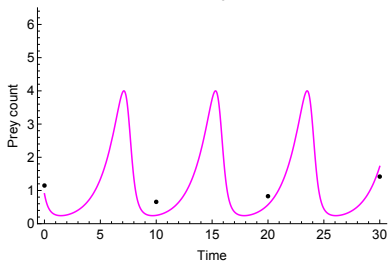
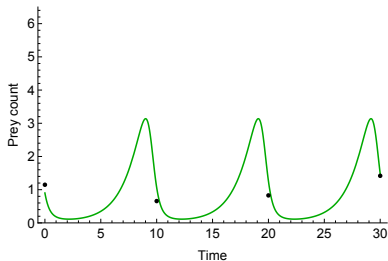
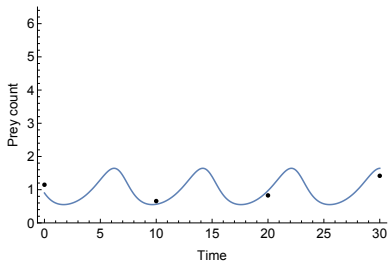
where $\tilde{x}(t)$ represents prey measurement at time t .

Measured prey series

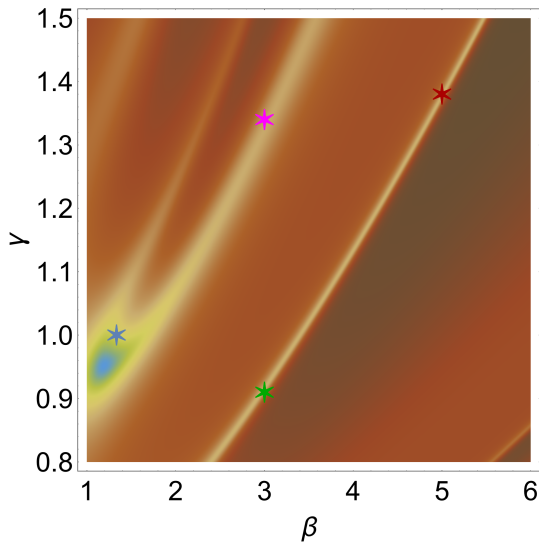
With $\beta = 4/3, \gamma = 1$.



Other explanations



Inverse distance surface



Lotka-Volterra inference problem: conclusions

sparse measurement + noise $\implies \sim$ poorly identified (15)

With the data to hand, it will be hard to estimate parameters with uncertainty. Solutions:

- Collect more data!
- Use pre-existing information to estimate parameters.