The (many) benefits of simulated data

Ben Lambert[1]
ben.c.lambert@gmail.com

[1]University 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 and 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:

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

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, $T(X)$:

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

where if $H_0 : \theta = 0$,

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

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

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

Statistical test size

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

Here,

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

is known as the size of a statistical test.

Statistical power

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

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

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 Y".

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_1s
3. You have determined and communicated the H_1s for which it doesn't work

$\Longrightarrow$ 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:

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

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

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

Or (say),

$$X_i = \rho X_{i-1} + \epsilon_i, \tag{9}$$

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_1s

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

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

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

Statistical significance is not practical significance

Suppose two treatments aimed at increasing personal income[1]:

- 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.
$\implies$ make decisions on practical utility based on changes to predictive power.

[1] From 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).

Example model: Lotka-Volterra

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

$$\frac{dx}{dt} = \alpha x - \beta xy \tag{11}$$

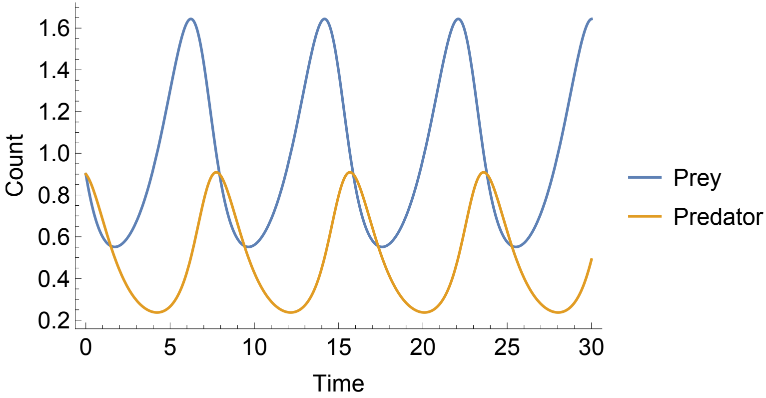
$$\frac{dy}{dt} = \delta xy - \gamma y \tag{12}$$

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)?

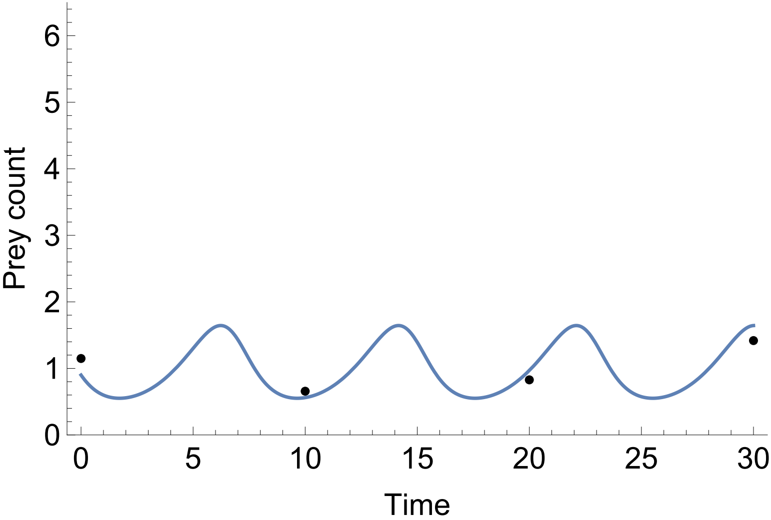
Answer: try inference for simulated data! Here, we assume same set of parameters as before and

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

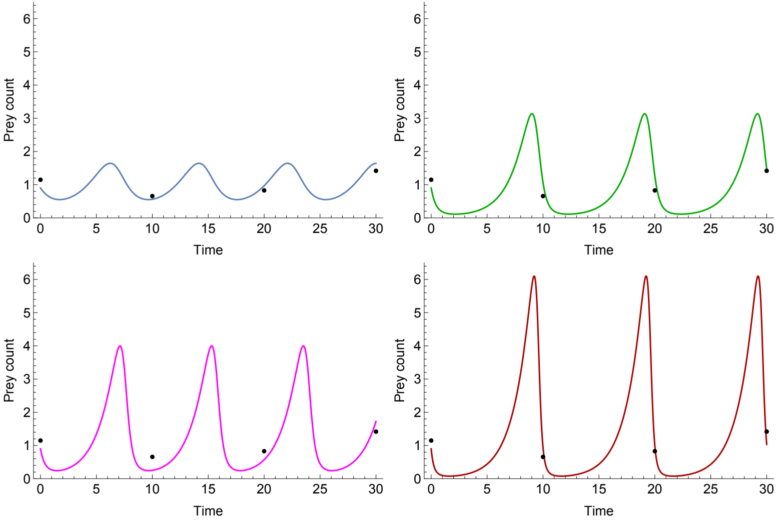
where $\tilde{x}(t)$ represents prey measurement at time t.

Measured prey series

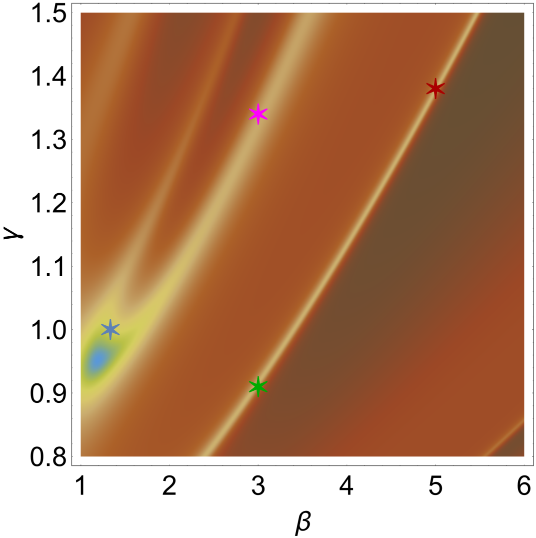
With $\beta = 4/3, \gamma = 1$.



Other explanations



Inverse distance surface



Lotka-Volterra inference problem: conclusions

$$\text{sparse measurement} + \text{noise} \implies \sim \text{poorly identified} \quad (15)$$

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.

Experimental design: the other side of the coin to inference

- Experiments typically aim to estimate certain quantities
- If we have choice about how to measure a system, we can affect the sampling distribution of our estimators
- Simulated data can be used to decide how best to measure

Using simulated data for experimental design

Repeat a number times at each experimental setup:

1. Simulate data using known parameters
2. Run inference on parameters
3. Compare true and estimated parameters

Note: for useful experimental design, the simulated data should be as near to what you expect as possible!

Parameter sensitivities: another tool for experimental design

Suppose we have a model with solution:

$$y(t) = f(t, \theta), \tag{16}$$

where t is time and θ is a parameter we wish to estimate. Suppose this then gets used to calculate a log-likelihood for inference:

$$\mathcal{L} = \sum_{t=t_1}^{t_T} \log p(y(t)|f(t, \theta)). \tag{17}$$

Parameter sensitivities: another tool for experimental design

The precision of our estimates depends on how sensitive the log-likelihood is to choice of θ. That is, on the magnitude of:

$$\frac{d\mathcal{L}}{d\theta}. \tag{18}$$

This, in turn, depends on:

$$\frac{df(t)}{d\theta}. \tag{19}$$

So assessing the sensitivities of our model at various points in time to the parameters can also be used to guide experimental design.

That's it!

Questions?