Special Topic Lecture: Implementing simulation studies

Author: Nicholas G Reich

This material is part of the **statsTeachR** *project*

What is simulation?

Definitions

- Broadly: "The technique of imitating the behaviour of some situation or process (whether economic, military, mechanical, etc.) by means of a suitably analogous situation or apparatus, esp. for the purpose of study or personnel training." (from the *OED*)
- In science: Creating a model that imitates a physical or biological process.
- In statistics: The generation of data from a model using rules of probability.

Simple examples of simulations

- Drawing pseudo-random numbers from a probability distribution (e.g. proposal distributions, ...).
- Generating data from a specified model (e.g. building a template dataset to test a method, calculating statistical power).
- Resampling existing data (e.g. permutation, bootstrap).

What simulations have you run?

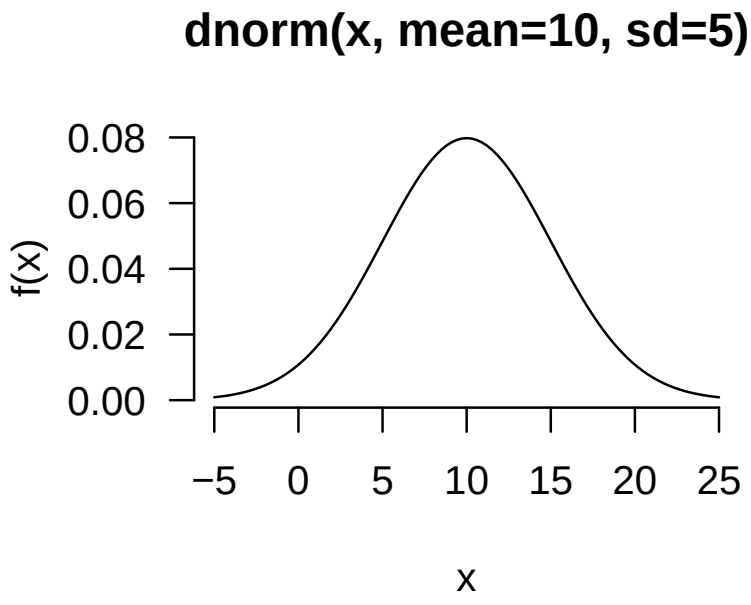
Random number generation is simulation

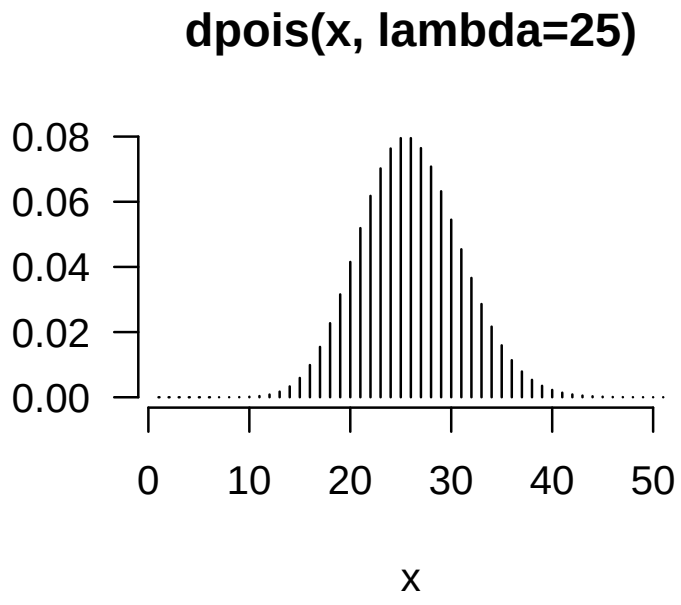
`rnorm()`, `rpois()`, etc...

Built-in functions for simulating from parametric distributions.

```
y <- rnorm(100, mean=10, sd=5)
(p <- rpois(5, lambda=25))

## [1] 19 23 37 23 27
```





Resampling data is simulation

`sample()`

Base R function for sampling data (with or without replacement).

```
p
```

```
## [1] 19 23 37 23 27
```

```
sample(p, replace=FALSE)
```

```
## [1] 19 23 37 23 27
```

```
sample(p, replace=TRUE)
```

```
## [1] 19 37 23 27 23
```

Generating data from a model is simulation

A Simple Linear Regression model

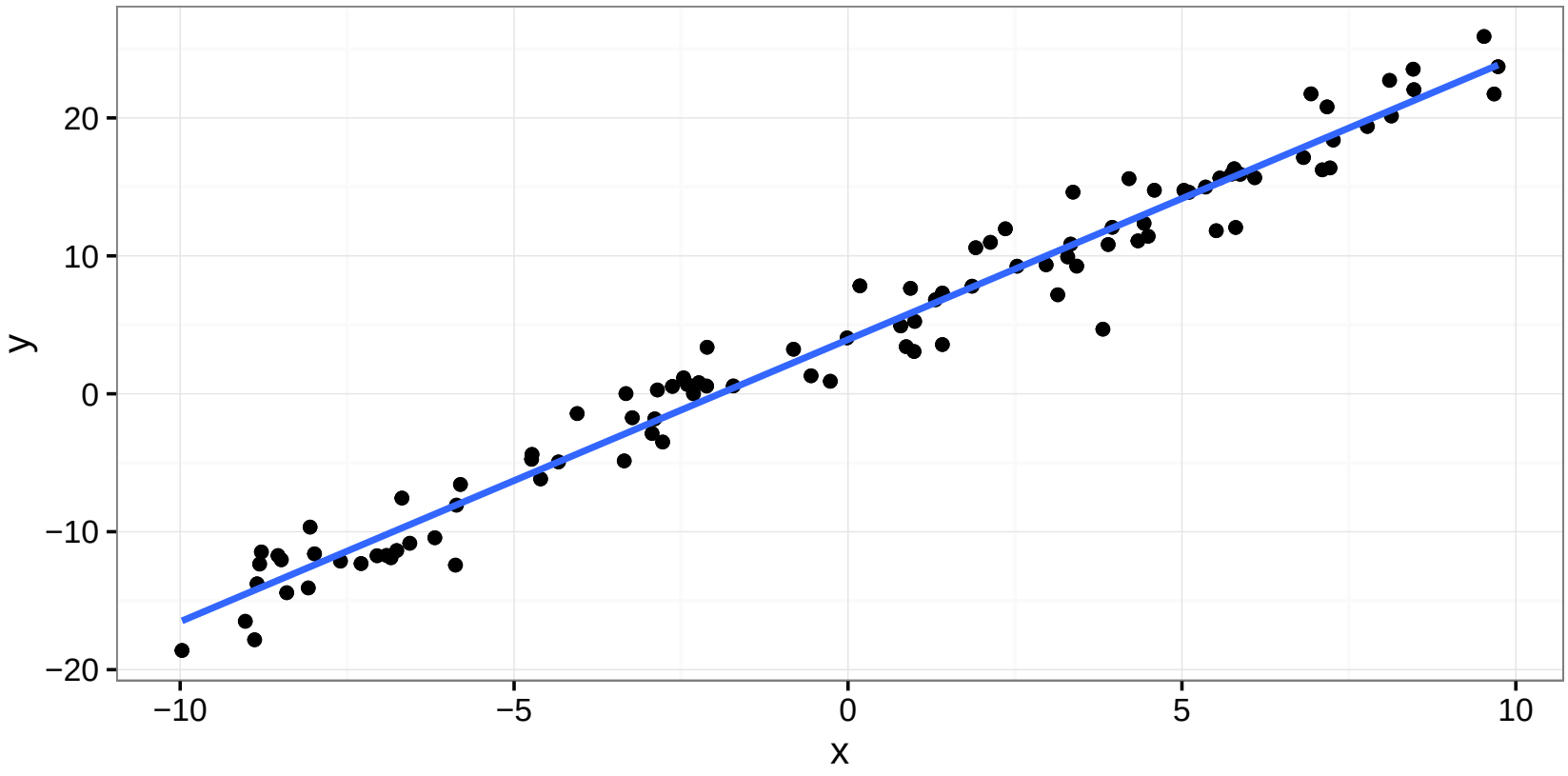
$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

What is needed to simulate data (i.e. Y_i) from this model?

- The X_i: fixed quantities.
- Error distribution: e.g. $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.
- Values for parameters: β_0, β_1, σ^2.

Generating data from $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

```
require(ggplot2)
n <- 100; b0=4; b1=2; sigma=2      ## define parameters
x <- runif(n, -10, 10)             ## fix the X's
eps <- rnorm(n, sd=sigma)          ## simulate the e_i's
y <- b0 + b1*x + eps               ## compute the y_i's
qplot(x, y) + geom_point() + geom_smooth(method="lm", se=FALSE)
```



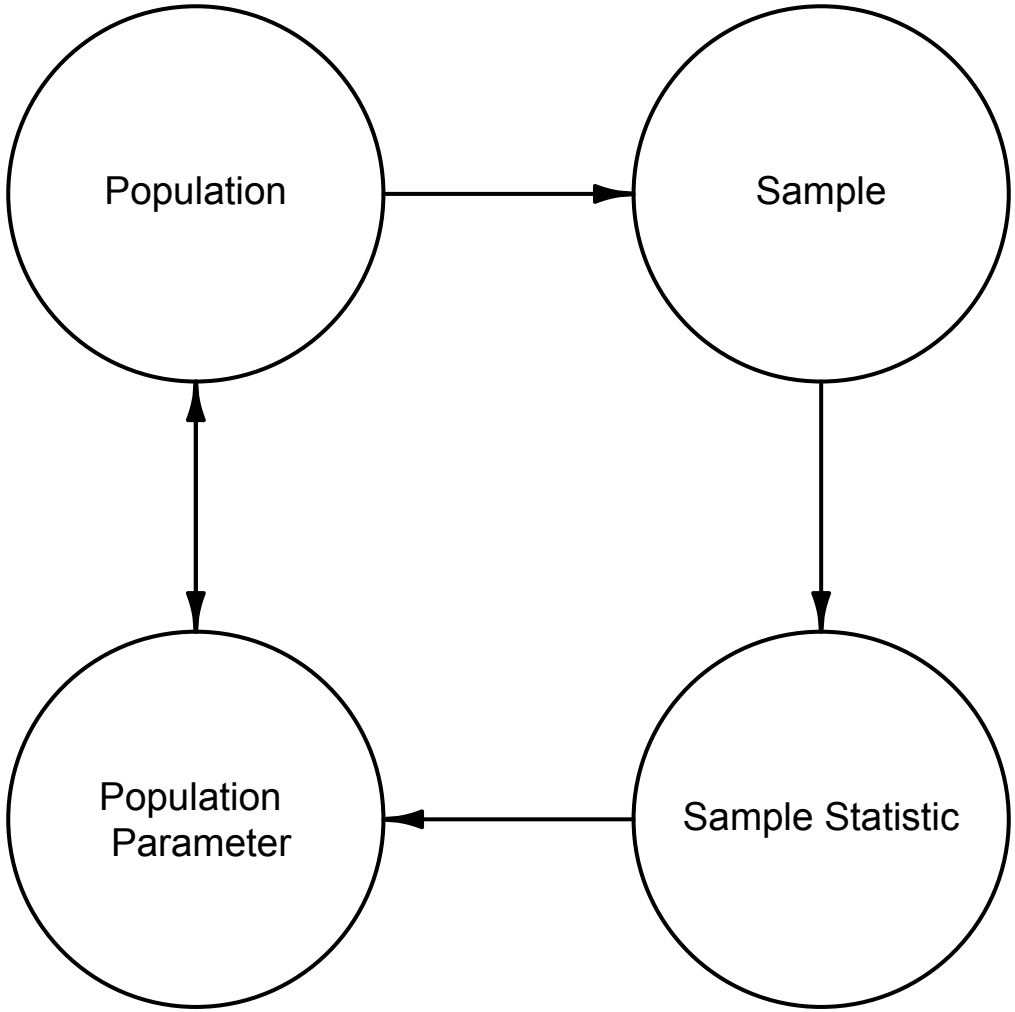
What is a simulation study?

Why run a simulation study?

- to evaluate whether your statistical method works!
- to determine how much variability you might reasonably expect from your estimates
- to calculate power for a study

Especially useful when data model and/or statistical method are complex, and do not have tidy theoretical results.

Circle of Life



What is a simulation study?

A statistician's laboratory

- control over the parameters of your data generating model
- systematic exploration of possible parameters
- careful evaluation of how one or more methods perform

Enables experimental answers to methodological questions

- Under a certain data model, which of a set of methods provides the best estimates of the true values?
- How much bias and variability can I expect from the method I am using on data that I want to analyze?
- MLR: How many predictors can I include before my methods become unreliable?
- MLR: What effect will correlation between predictors have on my estimates?

How to run a simulation study

Key steps

- Identify a data generating model and its associated parameters
- Define the question and scope: which parameters do you want to investigate? what ranges?
- Write code to run the analysis that is easily replicated (maybe write a function?)
- For each distinct set of parameters, generate and analyze data, storing the results. (Note: try to minimize operations within your loops, and consider running code that will parallelize easily)
- Summarize the results.

Recall: stepwise selection

E.g. Forward selection

- Start with "baseline" (usually intercept-only) model
- For every possible model that adds one term, evaluate the criterion you've settled on
- Choose the one with the best "score" (lowest AIC, smallest p-value)
- For every possible model that adds one term to the current model, evaluate your criterion
- Repeat until either adding a new term doesn't improve the model or all variables are included

It is with great ambivalence/trepidation that I reveal to you that stepwise selection can be easily implemented in R using, e.g. `stepAIC()` in the MASS package.

Let's design a simulation study about stepwise methods!

What specific method-performance measures could we evaluate?

Let's design a simulation study about stepwise methods!

What model/data features could we investigate?

Let's design a simulation study about stepwise methods!

Pick a low-risk, high-return combination of measures/features that you think will tell the best story.

Technical notes on simulations

- Decide how you will handle, report "failed" analyses prior to running the simulation. E.g. no convergence, etc...
- For reproducbility, consider setting a random seed using `set.seed()`. I like picking one from random.org
- Perform calculations that can be vectorized outside the simulation loop to save computational time. E.g. calculating bias.

Simulation study recap

- Simulation studies are key tools for evaluating different statistical methods.
- You will save time by planning your study carefully up front.
- We proposed some possible designs for a study on stepwise selection method performance. In small groups, you will work to answer these questions over the next few weeks.