Creating simulated data sets in R

Brad Duthie

8 July 2020

# Contents

**The ability to simulate data is a useful tool for better understanding statistical analyses and planning experimental designs. These notes illustrate how to simulate data using a variety of different functions in the R programming language, then discuss how data simulation can be used in research. These notes borrow heavily from a Stirling Coding Club session on** [**randomisation**](https://stirlingcodingclub.github.io/randomisation/randomisation_notes.html)**, and to a lesser extent from a session on** [**linear models**](https://stirlingcodingclub.github.io/linear_modelling/)**. After working through these notes, the reader should be able to simulate their own data sets and use them to explore data visualisations and statistical analysis. These notes are also available as** [**PDF**](https://stirlingcodingclub.github.io/linear_modelling/index.pdf) **and** [**DOCX**](https://stirlingcodingclub.github.io/linear_modelling/index.docx) **documents.**

* [Introduction: Simulating data](#intro)
* [Univariate random numbers](#univariate)
  + [Random uniform: runif](#runif)
  + [Random normal: rnorm](#rnorm)
  + [Random poisson: rpois](#rpois)
  + [Random binomial: rbinom](#rbinom)
* [Random sampling using sample](#sample)
  + [Sampling random numbers from a list](#sample_numbers)
  + [Sampling random characters from a list](#sample_characters)
* [Simulating data with known correlations](#mvtnorm)
* [Simulating a full data set](#fullsim)
* [Conclusions](#conclusions)
* [Literature Cited](#refs)

# Introduction: Simulating data

The ability generate simulated data is very useful in a lot of research contexts. Simulated data can be used to [better understand statistical methods](https://stirlingcodingclub.github.io/linear_modelling/#makeup), or in some cases to actually [run statistical analyses](https://stirlingcodingclub.github.io/randomisation/randomisation_notes.html#rand_m) (e.g., simulating a null distribution against which to compare a sample). Here I want to demonstrate how to simulate data in R. This can be accomplished with base R functions including [rnorm](#rnorm), [runif](#runif), [rbinom](#rbinom), [rpois](#rpois), or rgamma; all of these functions sample univariate data (i.e., one variable) from a specified distribution. The function [sample](#sample) can be used to sample elements from an R object with or without replacement. Using the [MASS](https://cran.r-project.org/web/packages/MASS/MASS.pdf) library, the [mvtnorm](#mvtnorm) function will sample multiple variables with a known correlation structure (i.e., we can tell R how variables should be correlated with one another) and normally distributed errors.

Below, I will first demonstrate how to use some common functions in R for simulating data. Then, I will illustrate how these simulated data might be used to better understand common statistical analyses and data visualisation.

# Univariate random numbers

Below, I introduce some base R functions that simulate (pseudo)random numbers from a given distribution. Note that most of what follows in this section is a recreation of a similar section in the notes for [randomisation](https://stirlingcodingclub.github.io/randomisation/randomisation_notes.html) analysis in R.

**Sampling from a uniform distribution**

Like the rnorm function, the runif function returns some number (n) of random numbers, but from a [uniform distribution](https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)) with a range from (min) to (max) such that , where . The default is to draw from a standard uniform distribution (i.e., and ).

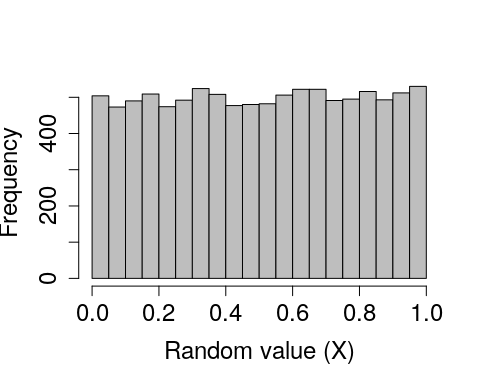
rand\_unifs\_10 <- runif(n = 10, min = 0, max = 1);

The above code stores a vector of ten numbers rand\_unifs\_10, shown below.

## [1] 0.9702199 0.6716896 0.0826678 0.8609107 0.1636970 0.1138159 0.6517036  
## [8] 0.2293410 0.9166176 0.9022468

We can visualise the standard uniform distribution that is generated by plotting a histogram of a very large number of values created using runif.

rand\_unifs\_10000 <- runif(n = 10000, min = 0, max = 1);  
hist(rand\_unifs\_10000, xlab = "Random value (X)", col = "grey",  
 main = "", cex.lab = 1.5, cex.axis = 1.5);



The random uniform distribution is special in some ways. The algorithm for generating random uniform numbers is the starting point for generating random numbers from other distributions using methods such as [rejection sampling](https://en.wikipedia.org/wiki/Rejection_sampling), [inverse transform sampling](https://en.wikipedia.org/wiki/Inverse_transform_sampling), or the [Box Muller](https://en.wikipedia.org/wiki/Box%E2%80%93Muller_transform) method (Box and Muller 1958).

**Sampling from a normal distribution**

The rnorm function returns some number (n) of randomly generated values given a set mean (; mean) and standard deviation (; sd), such that . The default is to draw from a standard [normal (a.k.a., “Gaussian”) distribution](https://en.wikipedia.org/wiki/Normal_distribution) (i.e., and ).

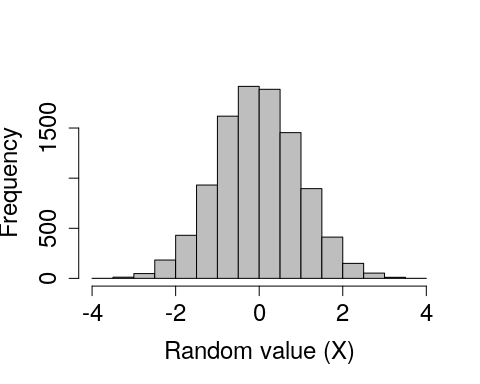
rand\_norms\_10 <- rnorm(n = 10, mean = 0, sd = 1);

The above code stores a vector of 10 numbers, shown below.

## [1] 0.9113831 -0.2814524 -0.1355369 0.2037147 1.2216920 0.2458136  
## [7] 0.1629585 -1.0100357 -0.9378795 -1.0157933

We can verify that a standard normal distribution is generated by plotting a histogram of a very large number of values created using rnorm.

rand\_norms\_10000 <- rnorm(n = 10000, mean = 0, sd = 1);  
hist(rand\_norms\_10000, xlab = "Random value (X)", col = "grey",  
 main = "", cex.lab = 1.5, cex.axis = 1.5);



Generating a histogram using data from a simulated distribution like this is often a useful way to visualise distributions, or to see how samples from the same distribution might vary. For example, if we wanted to compare the above distribution with a normal distribution that had a standard deviation of 2 instead of 1, then we could simply sample 10000 new values in rnorm with sd = 2 instead of sd = 1 and create a new histogram with hist. If we wanted to see what the distribution of sampled data might look like given a low sample size (e.g., 10), then we could repeat the process of sampling from rnorm(n = 10, mean = 0, sd = 1) multiple times and looking at the shape of the resulting histogram.

**Sampling from a poisson distribution**

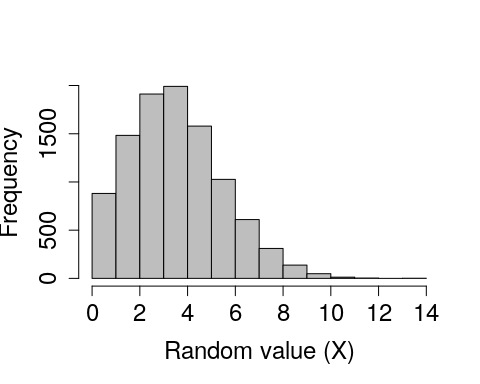
Many processes in biology can be described by a [Poisson distribution](https://en.wikipedia.org/wiki/Poisson_distribution). A Poisson process describes events happening with some given probability over an area of time or space such that , where the rate parameter is both the mean and variance of the Poisson distribution (note that by definition, , and values are always integers, as with count data). Sampling from a Poisson distribution can be done in R with rpois, which takes only two arguments specifying the number of values to be returned (n) and the rate parameter (lambda).

rand\_poissons <- rpois(n = 10, lambda = 1);  
print(rand\_poissons);

## [1] 2 1 2 2 1 0 2 1 1 1

There are no default values for rpois. We can plot a histogram of a large number of values to see the distribution when below.

rand\_poissons\_10000 <- rpois(n = 10000, lambda = 4);  
hist(rand\_poissons\_10000, xlab = "Random value (X)", col = "grey",  
 main = "", cex.lab = 1.5, cex.axis = 1.5);



**Sampling from a binomial distribution**

Sampling from a binomial distribution in R with rbinom is a bit more complex than using runif, rnorm, or rpois. Like those previous functions, the rbinom function returns some number (n) of random numbers, but the arguments and output can be slightly confusing at first. Recall that a [binomial distribution](https://en.wikipedia.org/wiki/Binomial_distribution) describes the number of ‘successes’ for some number of independent trials (). The rbinom function returns the number of successes after size trials, in which the probability of success in each trial is prob. For a concrete example, suppose we want to simulate the flipping of a fair coin 1000 times, and we want to know how many times that coin comes up heads (‘success’). We can do this with the following code.

coin\_flips <- rbinom(n = 1, size = 1000, prob = 0.5);  
print(coin\_flips);

## [1] 495

The above result shows that the coin came up heads 495 times. Note, however, the (required) argument n above. This allows the user to set the number of sequences to run. In other words, if we set n = 2, then this could simulate the flipping of a fair coin 1000 times once to see how many times heads comes up, then repeating the whole process a second time to see how many times heads comes up again (or, if it is more intuitive, the flipping of two separate fair coins 1000 times).

coin\_flips\_2 <- rbinom(n = 2, size = 1000, prob = 0.5);  
print(coin\_flips\_2);

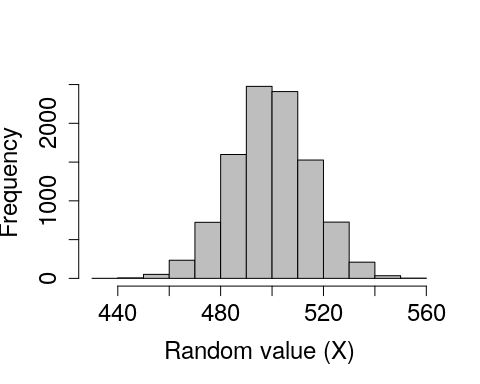
## [1] 521 491

In the above, a fair coin was flipped 1000 times and returned 521 heads, and then another fair coin was flipped 1000 times and returned 491 heads. As with the rnorm and runif functions, we can check to see what the distribution of the binomial function looks like if we repeat this process. Suppose, in other words, that we want to see the distribution of times heads comes up after 1000 flips. We can, for example, simulate the process of flipping 1000 times in a row with 10000 different coins using the code below.

coin\_flips\_10000 <- rbinom(n = 10000, size = 1000, prob = 0.5);

I have not printed the above coin\_flips\_10000 for obvious reasons, but we can use a histogram to look at the results.

hist(coin\_flips\_10000, xlab = "Random value (X)", col = "grey",  
 main = "", cex.lab = 1.5, cex.axis = 1.5);



As would be expected, most of the time ‘heads’ occurs around 500 times out of 1000, but usually the actual number will be a bit lower or higher due to chance. Note that if we want to simulate the results of individual flips in a single trial, we can do so as follows.

flips\_10 <- rbinom(n = 10, size = 1, prob = 0.5);

## [1] 1 1 0 0 1 1 1 1 0 1

In the above, there are n = 10 trials, but each trial consists of only a single coin flip (size = 1). But we can equally well interpret the results as a series of n coin flips that come up either heads (1) or tails (0). This latter interpretation can be especially useful to write code that randomly decides whether some event will happen (1) or not (0) with some probability prob.

# Random sampling using sample

Sometimes it is useful to sample a set of values from a vector or list. The R function sample is very flexible for sampling a subset of numbers or elements from some structure (x) in R according to some set probabilities (prob). Elements can be sampled from x some number of times (size) with or without replacement (replace), though an error will be returned if the size of the sample is larger than x but replace = FALSE (default).

**Sampling random numbers from a list**

To start out simple, suppose we want to ask R to pick a random number from one to ten with equal probability.

rand\_number\_1 <- sample(x = 1:10, size = 1);

The above code will set rand\_number\_1 to a randomly selected value, in this case 6. Because we have not specified a probability vector prob, the function assumes that every element in 1:10 is sampled with equal probability. We can increase the size of the sample to 10 below.

rand\_number\_10 <- sample(x = 1:10, size = 10);  
print(rand\_number\_10);

## [1] 6 3 1 9 5 4 10 8 7 2

Note that all numbers from 1 to 10 have been sampled, but in a random order. This is becaues the default is to sample with replacement, meaning that once a number has been sampled for the first element in rand\_number\_10, it is no longer available to be sampled again. To change this and allow for sampling with replacement, we can change the default.

rand\_number\_10\_r <- sample(x = 1:10, size = 10, replace = TRUE);  
print(rand\_number\_10\_r);

## [1] 5 10 7 8 7 4 5 5 7 10

Note that the numbers {5, 7, 10} are now repeated in the set of randomly sampled values above. We can also specify the probability of sampling each element, with the condition that these probabilities need to sum to 1. Below shows an example in which the numbers 1-5 are sampled with a probability of 0.05, while the numbers 6-10 are sampled with a probability of 0.15, thereby biasing sampling toward larger numbers.

prob\_vec <- c( rep(x = 0.05, times = 5), rep(x = 0.15, times = 5) );  
rand\_num\_bias <- sample(x = 1:10, size = 10, replace = TRUE, prob = prob\_vec);  
print(rand\_num\_bias);

## [1] 4 8 9 3 1 3 9 9 9 7

Note that rand\_num\_bias above contains more numbers from 6-10 than from 1-5.

**Sampling random characters from a list**

Sampling characters from a list of elements is no different than sampling numbers, but I am illustrating it separately because I find that I often sample characters for conceptually different reasons. For example, if I want to create a simulated data set that includes three different species, I might create a vector of species identities from which to sample.

species <- c("species\_A", "species\_B", "species\_C");

This gives three possible categories, which I can now use sample to draw from. Assume that I want to simulate the sampling of these three species, perhaps with species\_A being twice as common as species\_B and species\_C. I might use the following code to sample 12 times.

sp\_sample <- sample(x = species, size = 12, replace = TRUE,   
 prob = c(0.5, 0.25, 0.25)   
 );

Below are the values that get returned.

## [1] "species\_B" "species\_A" "species\_A" "species\_A" "species\_B" "species\_A"  
## [7] "species\_B" "species\_B" "species\_A" "species\_A" "species\_B" "species\_B"

# Simulating data with known correlations

We can generate variables and that have known correlations with with one another. The code below does this for two standard normal random variables with a sample size of 10000, such that the correlation between them is 0.3.

N <- 10000;  
rho <- 0.3;  
x1 <- rnorm(n = N, mean = 0, sd = 1);  
x2 <- (rho \* x1) + sqrt(1 - rho\*rho) \* rnorm(n = N, mean = 0, sd = 1);

Mathematically, these variables are generated by first simulating the sample (x1 above) from a standard normal distribution. Then, (x2 above) is calculated as below,

Where is a sample from a normal distribution with the same variance as . A simple call to the R function cor will confirm that the correlation does indeed equal rho.

cor(x1, x2);

## [1] 0.2976776

This is useful if we are only interested in two variables, but there is a much more efficient way to generate any number of variables with different variances and correlations to one another. To do this, we need to use the [MASS](https://cran.r-project.org/web/packages/MASS/MASS.pdf) library, which can be installed and loaded as below.

install.packages("MASS");  
library("MASS");

In the [MASS](https://cran.r-project.org/web/packages/MASS/MASS.pdf) library, the function mvrnorm can be used to generate any number of variables for a pre-specified covariance structure.

Suppose we want to simulate a data set of three measurements from a species of organisms. Measurement 1 () has a mean of 159.54 and variance of 12.68, measurement 2 () has a mean of 245.26 and variance of 30.39, and measurement 3 () has a mean of 25.52 and variance of 2.18. Below is a table summarising.

|  |  |  |
| --- | --- | --- |
| measurement | mean | variance |
| M1 | 159.54 | 12.68 |
| M2 | 245.26 | 30.39 |
| M3 | 25.52 | 2.18 |

Further, we want the covariance between and to equal 13.95, the covariance between and to equal 3.07, and the covariance between and to equal 4.7. We can put all of this information into a [covariance matrix](https://en.wikipedia.org/wiki/Covariance_matrix) with three rows and three columns. The diagonal of the matrix holds the variances of each variable, with the off-diagonals holding the covariances (note also that the variance of a variable is just the variable’s covariance with itself; e.g., ).

In R, we can create this matrix as follows.

matrix\_data <- c(12.68, 13.95, 3.07, 13.95, 30.39, 4.70, 3.07, 4.70, 2.18);  
cv\_mat <- matrix(data = matrix\_data, nrow = 3, ncol = 3, byrow = TRUE);  
rownames(cv\_mat) <- c("M1", "M2", "M3");  
colnames(cv\_mat) <- c("M1", "M2", "M3");

Here is what cv\_mat looks like (note that it is symmetrical along the diagonal).

## M1 M2 M3  
## M1 12.68 13.95 3.07  
## M2 13.95 30.39 4.70  
## M3 3.07 4.70 2.18

Now we can add the means to a vector in R.

mns <- c(159.54, 245.26, 25.52);

We are now ready to use the mvrnorm function in R to simulate some number n of sampled organisms with these three measurements. We use the mvrnorm arguments mu and Sigma to specify the vector of means and covariance matrix, respectively.

sim\_data <- mvrnorm(n = 40, mu = mns, Sigma = cv\_mat);

Here are the example data below.

|  |  |  |
| --- | --- | --- |
| M1 | M2 | M3 |
| 159.8460 | 247.8877 | 26.62498 |
| 168.0481 | 260.6712 | 26.75655 |
| 167.2071 | 254.7793 | 29.21997 |
| 159.1687 | 248.2836 | 24.01236 |
| 162.8051 | 243.4375 | 25.26723 |
| 160.1970 | 254.0904 | 24.69406 |
| 163.4046 | 246.2690 | 26.10888 |
| 159.3093 | 246.5744 | 26.10951 |
| 156.8984 | 238.9783 | 25.33972 |
| 155.9248 | 235.4744 | 24.25541 |
| 162.1055 | 244.5232 | 24.52365 |
| 160.2486 | 249.8029 | 26.57194 |
| 162.5742 | 242.1090 | 26.04947 |
| 158.4126 | 249.4451 | 26.61129 |
| 159.4849 | 243.0017 | 24.43700 |
| 165.2947 | 254.4368 | 27.62517 |
| 161.8141 | 246.9381 | 25.69734 |
| 161.8446 | 244.8448 | 25.81947 |
| 157.9735 | 241.0717 | 25.96450 |
| 158.1558 | 241.3307 | 25.23627 |
| 159.5748 | 241.3483 | 25.95585 |
| 163.9762 | 253.0840 | 24.77698 |
| 158.1150 | 247.6384 | 26.04184 |
| 164.4930 | 250.5311 | 25.44449 |
| 157.3527 | 248.2302 | 25.05499 |
| 156.4349 | 242.0303 | 26.49664 |
| 160.2729 | 240.6719 | 25.30131 |
| 167.0843 | 254.4179 | 27.59704 |
| 159.2069 | 243.7590 | 27.05021 |
| 159.2724 | 243.5871 | 24.44803 |
| 163.2690 | 247.4944 | 26.74340 |
| 159.0199 | 240.6654 | 24.34313 |
| 160.4763 | 243.7724 | 25.88892 |
| 159.4724 | 250.3505 | 25.11385 |
| 158.4743 | 240.4747 | 23.41712 |
| 154.8598 | 237.7489 | 26.45864 |
| 159.1277 | 244.9249 | 24.95962 |
| 159.1392 | 239.0690 | 24.26404 |
| 163.1686 | 255.8220 | 27.42379 |
| 159.3149 | 247.5697 | 27.37670 |

We can check to confirm that the mean values of each column are correct using apply.

apply(X = sim\_data, MARGIN = 2, FUN = mean);

## M1 M2 M3   
## 160.57057 246.17849 25.77703

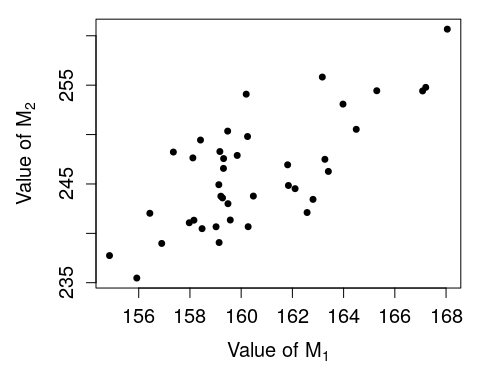
And we can check to confirm that the covariance structure of the data is correct using cov.

cov(sim\_data);

## M1 M2 M3  
## M1 9.500839 12.649878 1.693145  
## M2 12.649878 31.580784 3.334185  
## M3 1.693145 3.334185 1.441242

Note that the values are not exact, but should become closer to the specified values as increase the sample size n. We can visualse the data too; for example, we might look at the close correlation between and using a scatterplot, just as we would for data sampled from the field.

par(mar = c(5, 5, 1, 1));  
plot(x = sim\_data[,1], y = sim\_data[,2], pch = 20, cex = 1.25, cex.lab = 1.25,  
 cex.axis = 1.25, xlab = expression(paste("Value of ", M[1])),  
 ylab = expression(paste("Value of ", M[2])));



We could even run an ordination on these simulated data. For example, we could extract the principle components with prcomp, then plot the first two PCs to visualise these data. We might, for example, want to compare different methods of ordination using a data set with different, pre-specified properties (e.g., Minchin 1987). We might also want to use simulated data sets to investigate how different statistical tools perform. I show this in the next section, where I put a full data set together and run [linear models](https://stirlingcodingclub.github.io/linear_modelling/) on it.

# Simulating a full data set

Putting everything together, here I will create a data set of three different species from which three different measurements are taken. We can just call these measurements ‘length’, ‘width’, and ‘mass’. For simplicity, let us assume that these measurements always covary in the same way that we saw with (i.e., cv\_mat) above. But let’s also assume that we have three species with slightly different mean values. Below is the code that will build a new data set of samples with four columns: species, length, width, and mass.

N <- 20;  
matrix\_data <- c(12.68, 13.95, 3.07, 13.95, 30.39, 4.70, 3.07, 4.70, 2.18);  
cv\_mat <- matrix(data = matrix\_data, nrow = 3, ncol = 3, byrow = TRUE);  
mns\_1 <- c(159.54, 245.26, 25.52);  
sim\_data\_1 <- mvrnorm(n = N, mu = mns, Sigma = cv\_mat);  
colnames(sim\_data\_1) <- c("Length", "Width", "Mass");  
# Below, I bind a column for indicating 'species\_1' identity  
species <- rep(x = "species\_1", times = 20); # Repeats 20 times  
sp\_1 <- data.frame(species, sim\_data\_1);

Let us add one more data column. Suppose that we can also sample the number of offspring each organism has, and that the mean number of offspring that an organism has equals one tenth of the organism’s mass. To do this, we can use rpois, and take advantage of the fact that the argument lambda can be a vector rather than a single value. So to get the number of offspring for each organism based on its body mass, we can just insert the mass vector sp\_1$Mass times 0.1 for lambda.

offspring <- rpois(n = N, lambda = sp\_1$Mass \* 0.1);  
sp\_1 <- cbind(sp\_1, offspring);

I have also bound the offspring number to the data set sp\_1. Here is what it looks like below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| species | Length | Width | Mass | offspring |
| species\_1 | 162.0163 | 244.6611 | 27.20577 | 6 |
| species\_1 | 162.1346 | 243.4611 | 25.93084 | 7 |
| species\_1 | 157.8370 | 249.0522 | 25.45885 | 4 |
| species\_1 | 154.6387 | 237.0661 | 22.96693 | 3 |
| species\_1 | 162.0296 | 245.2809 | 25.68924 | 5 |
| species\_1 | 163.1852 | 244.8286 | 23.23753 | 2 |
| species\_1 | 159.6343 | 239.6246 | 24.46504 | 1 |
| species\_1 | 160.2166 | 247.8081 | 25.61839 | 4 |
| species\_1 | 158.3509 | 237.1585 | 23.41739 | 3 |
| species\_1 | 160.3624 | 246.6325 | 23.96519 | 1 |
| species\_1 | 162.9230 | 242.4909 | 26.75937 | 2 |
| species\_1 | 163.3319 | 252.7404 | 26.70315 | 4 |
| species\_1 | 168.5695 | 251.6960 | 28.21479 | 3 |
| species\_1 | 160.1818 | 246.7190 | 24.70195 | 1 |
| species\_1 | 162.4903 | 254.1485 | 27.14770 | 3 |
| species\_1 | 156.4505 | 239.8821 | 24.94645 | 4 |
| species\_1 | 155.0840 | 239.0621 | 24.66564 | 0 |
| species\_1 | 161.8352 | 249.3235 | 25.53353 | 1 |
| species\_1 | 166.0966 | 254.9537 | 26.70733 | 3 |
| species\_1 | 159.5090 | 240.1328 | 23.09964 | 3 |

To add two more species, let us repeat the process two more times, but change the expected mass just slightly each time. The code below does this, and puts everything together in a single data set.

# First making species 2  
mns\_2 <- c(159.54, 245.26, 25.52 + 3); # Add a bit  
sim\_data\_2 <- mvrnorm(n = N, mu = mns, Sigma = cv\_mat);  
colnames(sim\_data\_2) <- c("Length", "Width", "Mass");  
species <- rep(x = "species\_2", times = 20); # Repeats 20 times  
offspring <- rpois(n = N, lambda = sim\_data\_2[,3] \* 0.1);  
sp\_2 <- data.frame(species, sim\_data\_2, offspring);  
# Now make species 3  
mns\_3 <- c(159.54, 245.26, 25.52 + 4.5); # Add a bit more  
sim\_data\_3 <- mvrnorm(n = N, mu = mns, Sigma = cv\_mat);  
colnames(sim\_data\_3) <- c("Length", "Width", "Mass");  
species <- rep(x = "species\_3", times = 20); # Repeats 20 times  
offspring <- rpois(n = N, lambda = sim\_data\_3[,3] \* 0.1);  
sp\_3 <- data.frame(species, sim\_data\_3, offspring);  
# Bring it all together in one data set  
dat <- rbind(sp\_1, sp\_2, sp\_3);

Our full data set now looks like the below.

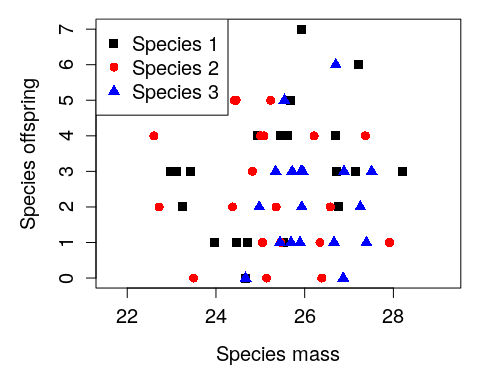
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| species | Length | Width | Mass | offspring |
| species\_1 | 162.0163 | 244.6611 | 27.20577 | 6 |
| species\_1 | 162.1346 | 243.4611 | 25.93084 | 7 |
| species\_1 | 157.8370 | 249.0522 | 25.45885 | 4 |
| species\_1 | 154.6387 | 237.0661 | 22.96693 | 3 |
| species\_1 | 162.0296 | 245.2809 | 25.68924 | 5 |
| species\_1 | 163.1852 | 244.8286 | 23.23753 | 2 |
| species\_1 | 159.6343 | 239.6246 | 24.46504 | 1 |
| species\_1 | 160.2166 | 247.8081 | 25.61839 | 4 |
| species\_1 | 158.3509 | 237.1585 | 23.41739 | 3 |
| species\_1 | 160.3624 | 246.6325 | 23.96519 | 1 |
| species\_1 | 162.9230 | 242.4909 | 26.75937 | 2 |
| species\_1 | 163.3319 | 252.7404 | 26.70315 | 4 |
| species\_1 | 168.5695 | 251.6960 | 28.21479 | 3 |
| species\_1 | 160.1818 | 246.7190 | 24.70195 | 1 |
| species\_1 | 162.4903 | 254.1485 | 27.14770 | 3 |
| species\_1 | 156.4505 | 239.8821 | 24.94645 | 4 |
| species\_1 | 155.0840 | 239.0621 | 24.66564 | 0 |
| species\_1 | 161.8352 | 249.3235 | 25.53353 | 1 |
| species\_1 | 166.0966 | 254.9537 | 26.70733 | 3 |
| species\_1 | 159.5090 | 240.1328 | 23.09964 | 3 |
| species\_2 | 159.1729 | 241.5189 | 25.00488 | 4 |
| species\_2 | 159.0781 | 249.1788 | 27.36802 | 4 |
| species\_2 | 154.5142 | 241.9229 | 22.71868 | 2 |
| species\_2 | 158.5578 | 243.1129 | 25.23427 | 5 |
| species\_2 | 167.9790 | 249.8829 | 26.34540 | 1 |
| species\_2 | 156.0300 | 242.6303 | 22.60039 | 4 |
| species\_2 | 159.8847 | 249.4163 | 26.38530 | 0 |
| species\_2 | 155.0206 | 239.8918 | 24.37642 | 2 |
| species\_2 | 160.9491 | 243.5860 | 24.82099 | 3 |
| species\_2 | 163.9520 | 247.1705 | 27.91111 | 1 |
| species\_2 | 159.1194 | 241.7315 | 26.57770 | 2 |
| species\_2 | 164.0003 | 249.8026 | 25.08043 | 4 |
| species\_2 | 153.8578 | 247.9412 | 25.45756 | 1 |
| species\_2 | 158.9797 | 242.4058 | 24.41459 | 5 |
| species\_2 | 160.2418 | 248.1337 | 25.14239 | 0 |
| species\_2 | 159.3784 | 240.1066 | 26.21218 | 4 |
| species\_2 | 155.6436 | 243.7619 | 24.45449 | 5 |
| species\_2 | 161.7986 | 240.0830 | 25.04731 | 1 |
| species\_2 | 158.2877 | 239.9688 | 25.35600 | 2 |
| species\_2 | 152.5190 | 240.4154 | 23.49561 | 0 |
| species\_3 | 155.4456 | 239.1643 | 24.97365 | 2 |
| species\_3 | 162.9985 | 246.6504 | 25.71843 | 3 |
| species\_3 | 163.8482 | 241.9071 | 25.89463 | 1 |
| species\_3 | 161.6252 | 239.1157 | 25.93643 | 2 |
| species\_3 | 165.1217 | 244.9714 | 26.86908 | 0 |
| species\_3 | 165.1981 | 252.4080 | 25.95565 | 3 |
| species\_3 | 154.3892 | 242.0033 | 24.66740 | 0 |
| species\_3 | 158.5160 | 237.9390 | 23.71664 | 5 |
| species\_3 | 158.7108 | 239.9081 | 25.54282 | 5 |
| species\_3 | 162.7083 | 244.8092 | 26.70040 | 6 |
| species\_3 | 157.6075 | 242.9777 | 27.25615 | 2 |
| species\_3 | 163.1200 | 242.4084 | 27.51086 | 3 |
| species\_3 | 159.2623 | 246.0800 | 27.39178 | 1 |
| species\_3 | 162.3009 | 247.9707 | 26.66279 | 1 |
| species\_3 | 160.5197 | 250.3616 | 25.44475 | 1 |
| species\_3 | 158.1010 | 245.3060 | 25.91898 | 3 |
| species\_3 | 160.7495 | 249.0713 | 25.34665 | 3 |
| species\_3 | 161.8069 | 252.9446 | 26.88912 | 3 |
| species\_3 | 154.4150 | 235.8856 | 25.69330 | 1 |
| species\_3 | 150.3532 | 239.0640 | 22.68467 | 5 |

To summarise, we now have a simulated data set of measurements from three different species, all of which have known variances and covariances of length, width, and mass. Each species has a slightly different mean mass, and for all species, each unit of mass increases the expected number of offspring by 0.1. Because we know these properties of the data for certain, we can start asking questions that might be useful to know about our data analysis. For example, given this covariance structure and these small differences in mass, is a sample size of 20 really enough to even get a significant difference among species masses using an ANOVA? What if we tried to test for differences among masses using some sort of [randomisation approach](https://stirlingcodingclub.github.io/randomisation/randomisation_notes.html#rand_h) Instead? Would this give us more or less power? Let us run an ANOVA to see if the difference between group means (which we know exists) is recovered.

aov\_result <- aov(Mass ~ species, data = dat);  
summary(aov\_result);

## Df Sum Sq Mean Sq F value Pr(>F)  
## species 2 4.6 2.299 1.22 0.303  
## Residuals 57 107.4 1.884

It appears not! What about the relationship between body mass and offspring production that we know exists? Below is a scatterplot of the data for the three different species.



This looks like there might be a postive relationship. We can use a generalised linear model to test it with species as a random effect, as we might do if these were data sampled from the field.

library(lme4);

## Loading required package: Matrix

mod <- glmer(offspring ~ Mass + (1 | species), data = dat, family = "poisson");

## boundary (singular) fit: see ?isSingular

summary(mod);

## Generalized linear mixed model fit by maximum likelihood (Laplace  
## Approximation) [glmerMod]  
## Family: poisson ( log )  
## Formula: offspring ~ Mass + (1 | species)  
## Data: dat  
##   
## AIC BIC logLik deviance df.resid   
## 238.8 245.0 -116.4 232.8 57   
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.6486 -1.0120 0.1671 0.8111 2.6677   
##   
## Random effects:  
## Groups Name Variance Std.Dev.   
## species (Intercept) 2.404e-17 4.903e-09  
## Number of obs: 60, groups: species, 3  
##   
## Fixed effects:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 1.229002 1.471825 0.835 0.404  
## Mass -0.009754 0.057782 -0.169 0.866  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## Mass -0.999  
## convergence code: 0  
## boundary (singular) fit: see ?isSingular

There does not appear to be any effect here either! To get one, it appears that we will need to simulate a larger data set (or a bigger effect size – or just get lucky when re-simulating a new data set).

Note that I have run a linear model that might be reasonable given the structure of our data. But the advantage of working with simulated data and knowing for certain what the relationship is between the underlying variables is that we can explore different statistical techniques. For example, we know that our response variable offspring is count data, so we are supposed to specify a Poisson error structure using the family = "poisson" argument above, right? But what would happen if we just used a normal error structure anyway? Would this really be so bad? Now is the opportunity to test because we *know* what the correct answer is supposed to be! Trying statistical methods that are normally ill-advised can actually be useful here, as it can help us see for ourselves when a technique is bad – or perhaps when it really is not (e.g., Ives 2015).

# Conclusions

Simulating data can be a powerful tool for learning and investigating different statistical analyses. The main benefits of using simulated data are flexibility and certainty. Simulation gives us the flexibility to explore any number of hypotheticals, including different sample sizes, effect sizes, relationships between variables, and error distributions. It also works from a point of certainty; we know what the real relationship is between variables, and what the actual effect sizes are because we define them when generating random samples. So if we want to better understand what would happen if we were unable to sample an important variable in our system, or if we were to use a biased estimator, or if we we were to violate key model assumptions, simulated data is a very useful tool.

# Literature cited

Box, G E P, and Mervin E Muller. 1958. “A note on the generation of random normal deviates.” *The Annals of Mathematical Statistics* 29 (2): 610–11. <https://doi.org/10.1214/aoms/1177706645>.

Ives, Anthony R. 2015. “For testing the significance of regression coefficients, go ahead and log-transform count data.” *Methods in Ecology and Evolution* 6: 828–35. <https://doi.org/10.1111/2041-210X.12386>.

Minchin, Peter R. 1987. “An evaluation of the relative robustness of techniques for ecological ordination.” *Vegetatio* 69 (1-3): 89–107. <https://doi.org/10.1007/BF00038690>.