Short case study 1:
Is a sample size of five stormy islands enough?

A simple data simulation example (Step 2)
From: A four-step simulation-based workflow for ecological analysis and science

Simulating data and then fitting your model to those simulated data is a powerful way to interrogate both our statistical and scientific models, but learning how to do it means starting simple.

Here we imagine a very simple example where you are interested in whether a storm that hit several tropical islands, destroying vegetation on the islands, may have led lizards to evolve shorter toes—since they may no longer need longer toes to run up vegetation. You plan to sample one lizard per island (they're hard to catch and you have limited time), on 10 islands—including 5 of them that were not hit by the storm.

You could build a very simple linear regression model to simulate these data:

```
set.seed(7799)

basetoesize <- 5 # intercept of the model (mean toe length across islands)
stormeffect <- -0.5 # reduction of toe length caused by the storm
sigmay <- 0.5 # measurement error
reps <- 5
x <- c(rep(0, reps), rep(1, reps)) # 5 devastated islands, 5 spared islands

y <- basetoesize + stormeffect*x + rnorm(length(x), 0, sigmay)
```

And then you could fit your statistical model to it and see the results:

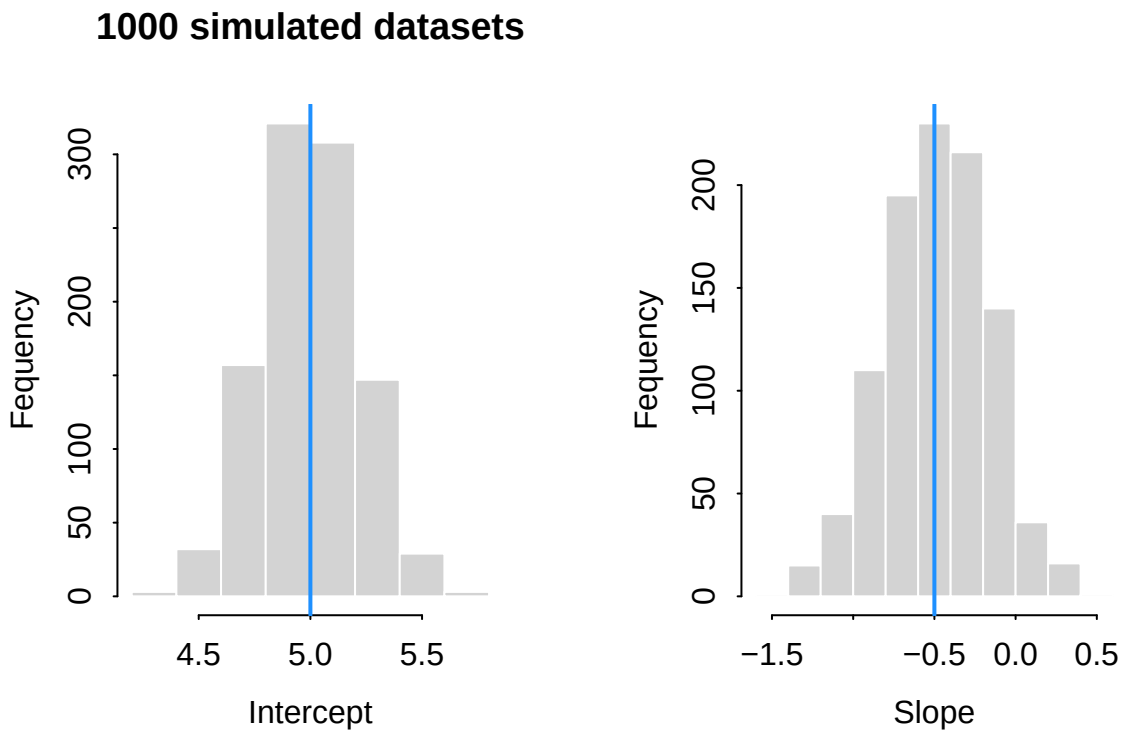
```
summary(lm(y~x))
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53514 -0.17525 -0.07043  0.32105  0.49173
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9974     0.1760  28.397 2.56e-09 ***
## x            -0.3676     0.2489  -1.477    0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3935 on 8 degrees of freedom
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.116
## F-statistic: 2.181 on 1 and 8 DF,  p-value: 0.1779
```

Your model is returning your parameters pretty well, but this is only one draw of all the possible datasets you might simulate. You can repeat this process 1000 times, however:

```
df <- data.frame(intercept=numeric(), slope=numeric())
for(i in c(1:1000)){
    y <- basetoesize + stormeffect*x + rnorm(length(x), 0, sigmay)
    df[i,] <- coef(lm(y~x))
}

par(mfrow=c(1,2), mgp=c(2, 0.5, 0), tck=-0.01)
hist(df$intercept, bty="l", ylab="Fequency", xlab="Intercept",
    main="1000 simulated datasets", border = 'white')
abline(v=basetoesize, col="dodgerblue", lwd=2)
hist(df$slope, bty="l", ylab="Fequency", xlab="Slope",
     main="", border = 'white')
abline(v=stormeffect, col="dodgerblue", lwd=2)
```




```
nrow(subset(df, slope> 0))
```

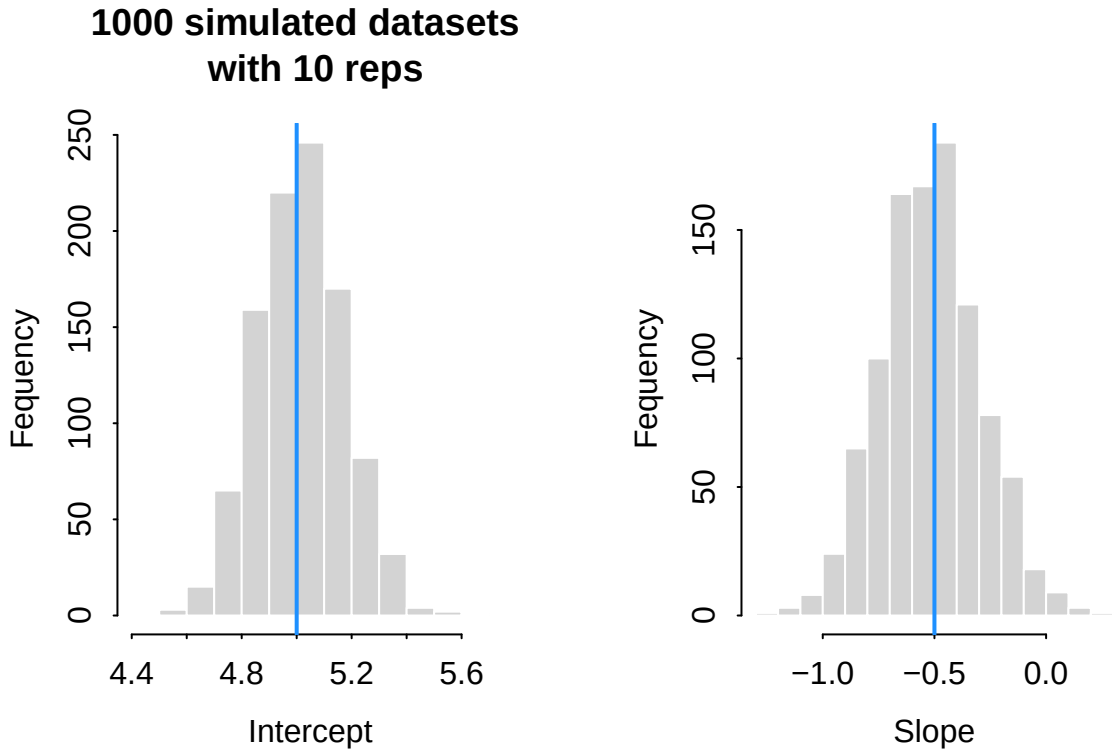
```
## [1] 53
```

The statistical model gets close to your assigned slope and intercept on average but with a lot of variation. About 5% of time you see no effect or a positive effect even, when you assigned a negative effect of the storm to toe length.

What would happen if you somehow could sample more and get 10 reps of each? You can easily test this possibility.

```
reps <- 10
x <- c(rep(0, reps), rep(1, reps))
df <- data.frame(intercept=numeric(), slope=numeric())
for(i in c(1:1000)){
    y <- basetoesize + stormeffect*x + rnorm(length(x), 0, sigmay)
    df[i,] <- coef(lm(y~x))
}
```

```
par(mfrow=c(1,2), mgp=c(2, 0.5, 0), tck=-0.01)
hist(df$intercept, bty="l", ylab="Fequency", xlab="Intercept",
    main="1000 simulated datasets \n with 10 reps", border = 'white')
abline(v=basetoesize, col="dodgerblue", lwd=2)
hist(df$slope, bty="l", ylab="Fequency", xlab="Slope",
     main="", border = 'white')
abline(v=stormeffect, col="dodgerblue", lwd=2)
```




```
nrow(subset(df, slope> 0))
```

```
## [1] 13
```

Now only about 1% of time you see no effect or a positive effect.

You can equally adjust other parameters to see their effect on returning your parameter estimates, or see how often your p-value is significant if that is of interest. You might also start considering other experimental designs such as sampling multiple lizards per island and adjust your simulated data and model to reflect this. This might get you to thinking about the evolution of toe length on each island, and lead you to model how that happens, which might in turn alter your simulated data and statistical model – and it would mean you've gone from thinking about the problem as a linear regression to something closer to a biological model, all before you set off for those tropical islands!