Simulating data in R

stirlingcodingclub.github.io/simulating_data

Stirling Coding Club

14 December 2022

Why simulate data?

Simulating data uses generating random data sets with known properties using code (or some other method). This can be useful in a lot of contexts.

- Better understand statistical methods
- Plan ahead for actual data collection
- Visualise data sets and distributions
- Run some statistical analyses (randomisation)

How can data be simulated in R?

Random data with different properties can be generated in R using several base R functions.

- ‘runif‘ generates random values from a uniform distribution.

How can data be simulated in R?

Random data with different properties can be generated in R using several base R functions.

- ‘runif‘ generates random values from a uniform distribution.
- ‘rnorm‘ generates random values from a normal distribution.

How can data be simulated in R?

Random data with different properties can be generated in R using several base R functions.

- ‘runif‘ generates random values from a uniform distribution.
- ‘rnorm‘ generates random values from a normal distribution.
- ‘rpois‘ generates random values from a poisson distribution.

How can data be simulated in R?

Random data with different properties can be generated in R using several base R functions.

- ‘runif‘ generates random values from a uniform distribution.
- ‘rnorm‘ generates random values from a normal distribution.
- ‘rpois‘ generates random values from a poisson distribution.
- ‘rbinom‘ generates random values from a binomial distribution.

How can data be simulated in R?

Random data with different properties can be generated in R using several base R functions.

- ‘runif‘ generates random values from a uniform distribution.
- ‘rnorm‘ generates random values from a normal distribution.
- ‘rpois‘ generates random values from a poisson distribution.
- ‘rbinom‘ generates random values from a binomial distribution.
- ‘sample‘ samples values from any given vector with or without replacement.

Other R packages, such as the MASS library, can simulate full data sets with pre-defined correlation stuctures.

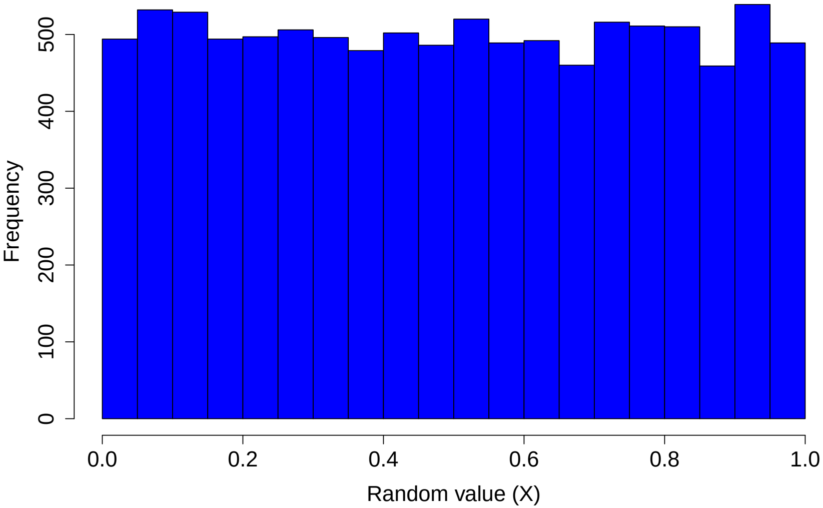
The runif function in R

```
rand_unifs <- runif(n = 10000, min = 0, max = 1);
```

```
##  [1] 0.71739297 0.75035000 0.56013160 0.24498652 0.68830
##  [7] 0.67665735 0.20360831 0.52135276 0.48068556 0.55222
## [13] 0.66911110 0.85938387 0.58939530 0.73200866 0.72432
## [19] 0.12922075 0.57770871 0.98332661 0.28403160 0.72911
## [25] 0.96192536 0.13342066 0.90172250 0.54101509 0.49679
## [31] 0.85583746 0.15100820 0.07010486 0.83421067 0.94163
## [37] 0.94881335 0.89494764 0.85449798 0.32408480
```

The `runif` function in R

```
rand_unifs <- runif(n = 10000, min = 0, max = 1);
```



The runif function in R

```
int verify_seed(int x){
  x=abs(x) % 30000;
  return(++x);
 } /* Easy way of getting seeds */

double as183(int seeds[]){
  double unidev; /* Code below verifies the 3 seeds */
  seeds[0] = verify_seed(seeds[0]);
  seeds[1] = verify_seed(seeds[1]);
  seeds[2] = verify_seed(seeds[2]);
  /* Code below gets a decimal to be added to unidev */
  seeds[0] = (171 * seeds[0]) % 30269;
  seeds[1] = (172 * seeds[1]) % 30307;
  seeds[2] = (170 * seeds[2]) % 30323;
  /* unidev gets a random uniform number between zero and one */
  unidev = seeds[0]/30269.0 + seeds[1]/30307.0 + seeds[2]/30323.
  /* Return just the decimal, subtract integer of unidev */
  return(unidev - (int)unidev);
} /* We now have one random uniform number */
```

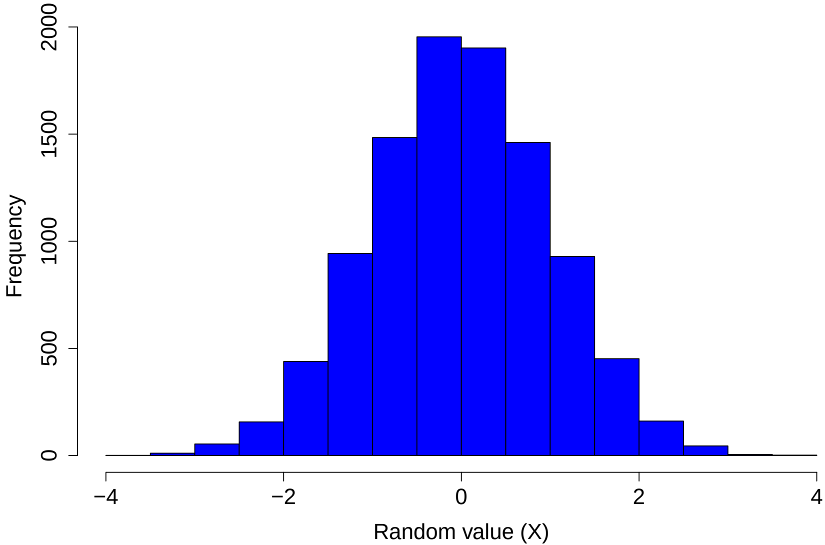
The rnorm function in R

```r
rand_rnorms <- rnorm(n = 10000, mean = 0, sd = 1);
```

```
##  [1]  0.94726152  0.68356269  0.49704858  0.44813645  0.
##  [7]  0.24372762  1.53200867  0.51274151  0.59188341  1.
## [13]  0.98859434 -1.32302537  0.12578028  0.78844279  0.
## [19] -0.60932700 -0.38531844 -1.07812230  0.44641907  0.
## [25] -1.60425962  0.98398122  2.03062832 -1.55893601 -1.
## [31]  0.10309797  0.62000705  0.41811461  1.26991674 -0.
## [37] -0.83093977 -0.31985344  0.93327407  0.19218058
```

The rnorm function in R

```
rand_rnorms <- rnorm(n = 10000, mean = 0, sd = 1);
```



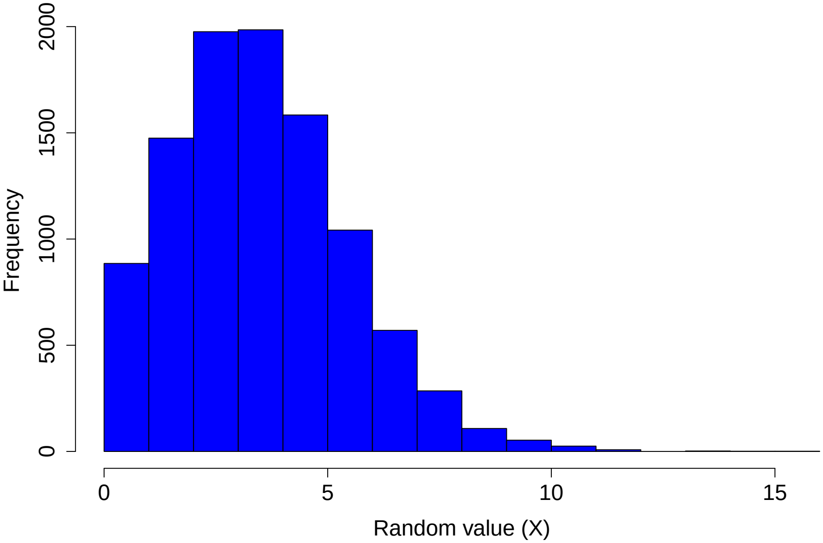
The rpois function in R

```r
rand_rpois <- rpois(n = 10000, lambda = 4);
```

```
##  [1]  0  7  3  3  4  5  5  2  4  6  6  0  6  5  7  7  6
## [26]  3  5  7  3  3  3  3  3  7  3  4  5  3  9  4
```

The rpois function in R

```
rand_rpois <- rpois(n = 10000, lambda = 4);
```



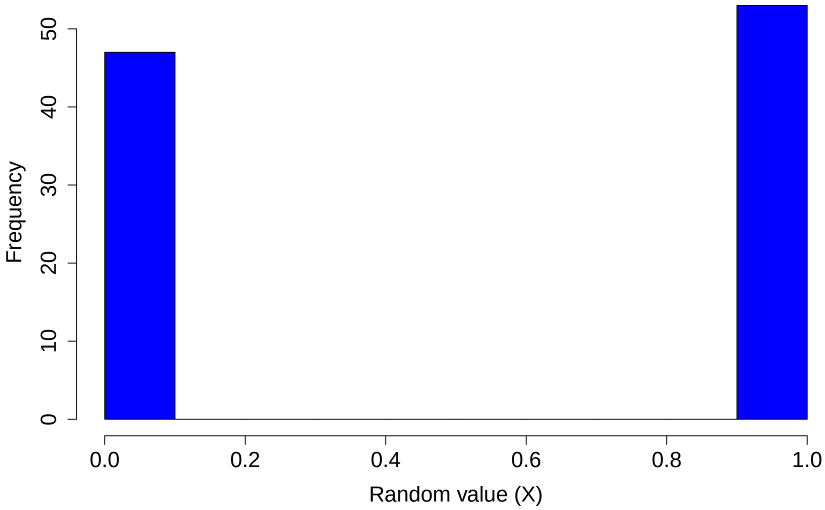
The rbinom function in R

```r
rand_rbinom <- rbinom(n = 100, size = 1, prob = 0.5);
```

```
##  [1] 0 0 1 0 1 1 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0
## [39] 1 1
```

The rpois function in R

```r
rand_rbinom <- rbinom(n = 100, size = 1, prob = 0.5);
```



Using sample in R

Create a vector of numbers from which to sample.

```
my_sample_vec <- 1:10;
```

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

Using `sample` in R

Create a vector of numbers from which to sample.

```
my_sample_vec <- 1:10;
```

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

Use `sample` to randomly sample numbers from `my_sample_vec`

```
my_sample <- sample(x = my_sample_vec, size = 4);
```

```
## [1] 7 6 5 8
```

Using sample in R

Can sample with or without replacement.

```
sample_no_replace <- sample(x = my_sample_vec,
                            size = 10, replace = FALSE);
```

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

Using sample in R

Can sample with or without replacement.

```r
sample_no_replace <- sample(x = my_sample_vec,
                            size = 10, replace = FALSE);
```

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

```r
sample_replace <- sample(x = my_sample_vec,
                         size = 10, replace = TRUE);
```

```
##  [1]  6 10  6  9  6 10  5  5  5  8
```

Using sample in R

Can also change the probabilities of being sampled

```r
# Vector values must sum to 1
pr_vector  <- c(0, 0, 0, 0, 0,
                0.2, 0.2, 0.2,
                0.2, 0.2);
new_sample <- sample(x = 1:10, size = 10,
                     replace = TRUE,
                     prob = pr_vector);
```

```
##  [1] 10  8  6  7 10  8  9 10  6  7
```

Using sample in R

Can also sample strings instead of numbers

```r
species   <- c("species_A", "species_B", "species_C");
sp_sample <- sample(x = species, size = 12,
                    replace = TRUE,
                    prob = c(0.5, 0.25, 0.25)
                    );
```

```
##  [1] "species_B" "species_A" "species_C" "species_A" "sp
##  [7] "species_B" "species_A" "species_A" "species_B" "sp
```

Building a simple simulated dataset

```
N          <- 12;
species    <- c("species_A", "species_B");
sp_sample  <- sample(x = species,
                    size = N, replace = TRUE);
sp_mass    <- rnorm(n = N, mean = 100, sd = 4);
for(i in 1:N){
  if(sp_sample[i] == "species_A"){
    sp_mass[i] <- sp_mass[i] + rnorm(n = 1,
                                     mean = 4, sd = 1);
  }
}
sim_data  <- data.frame(sp_sample, sp_mass);
```

Building a simple simulated dataset

sp_sample	sp_mass
species_B	97.58065
species_B	96.96083
species_B	104.09267
species_A	105.05708
species_A	102.81067
species_A	106.13235
species_A	104.61697
species_A	107.89034
species_B	101.01198
species_B	96.17857
species_B	100.64885
species_B	103.06411

Building a simple simulated dataset

```r
t.test(sp_mass ~ sp_sample, data = sim_data);
```

```
##
##  Welch Two Sample t-test
##
## data:  sp_mass by sp_sample
## t = 3.7314, df = 9.8585, p-value = 0.003999
## alternative hypothesis: true difference in means betwee
## 95 percent confidence interval:
##  2.156122 8.578935
## sample estimates:
## mean in group species_A mean in group species_B
##               105.30148                99.93395
```

Building a simple simulated dataset

```r
N         <- 120;
species   <- c("species_A", "species_B");
sp_sample <- sample(x = species, size = N,
                    replace = TRUE);
sp_mass   <- rnorm(n = N, mean = 100, sd = 4);
for(i in 1:N){
  if(sp_sample[i] == "species_A"){
    sp_mass[i] <- sp_mass[i] + rnorm(n = 1,
                                     mean = 4, sd = 1);
  }
}
sim_data  <- data.frame(sp_sample, sp_mass);
```

Building a simple simulated dataset

```r
t.test(sp_mass ~ sp_sample, data = sim_data);
```

```
##
##  Welch Two Sample t-test
##
## data:  sp_mass by sp_sample
## t = 4.7219, df = 92.007, p-value = 8.341e-06
## alternative hypothesis: true difference in means betwee
## 95 percent confidence interval:
##  2.087618 5.118701
## sample estimates:
## mean in group species_A mean in group species_B
##                104.0913                100.4882
```

Setting a seed gives the same numbers

Try it once

```r
set.seed(10);
rnorm(n = 10);
```

```
##  [1]  0.01874617 -0.18425254 -1.37133055 -0.59916772  0.
##  [7] -1.20807618 -0.36367602 -1.62667268 -0.25647839
```

Try it again

```r
set.seed(10);
rnorm(n = 10);
```

```
##  [1]  0.01874617 -0.18425254 -1.37133055 -0.59916772  0.
##  [7] -1.20807618 -0.36367602 -1.62667268 -0.25647839
```